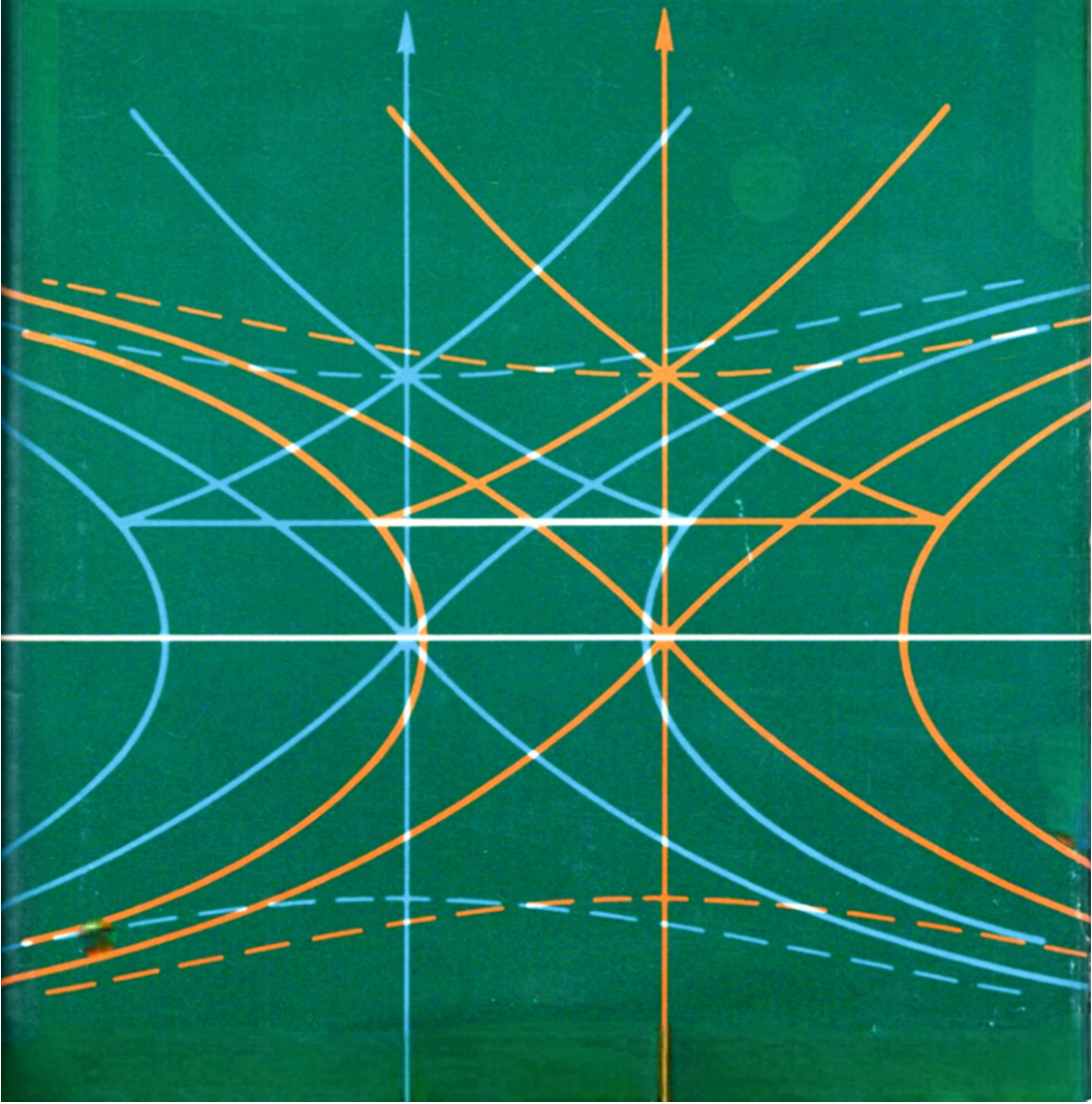
Relativistic Astrophysics

Ya. B. Zeldovich and
I. D. Novikov

Volume 2

The Structure and Evolution of The Universe



RELATIVISTIC ASTROPHYSICS

(In Two Volumes)

Volume 2

THE STRUCTURE AND EVOLUTION OF THE UNIVERSE

To the memory of my father and to my mother.

Gary Steigman

The University of Chicago Press acknowledges with gratitude the efforts of Kip S. Thorne, the William R. Kenan, Jr., Professor and professor of theoretical physics, California Institute of Technology, on behalf of this book. During the twelve years required for its translation, updating, and editing, Professor Thorne offered assistance, advice, and encouragement. When problems arose, time and again he set aside his own work to solve them. He also enlisted the help of many others, including JoAnn Boyd Anderson, who transliterated and organized the references; Mike Wilson, who edited chapter 16; Bernard Carr, who edited chapter 17; Douglas Macdonald, who edited the last half of chapter 24; and Ian H. Redmount, who read the page proofs of the entire book and made final checks of the equations against the text. Professor Thorne compiled the index.

RELATIVISTIC ASTROPHYSICS

VOLUME 2

The Structure and Evolution of the Universe

YA. B. ZEL'DOVICH AND I. D. NOVIKOV

Revised and Enlarged from the Original Russian Edition
by the Authors and the Editor

Translated by LESLIE FISHBONE

EDITED BY
GARY STEIGMAN

THE UNIVERSITY OF CHICAGO PRESS
Chicago and London

Ya. B. Zel'dovich is a member of the USSR Academy of Sciences, he is leader of an internationally renowned research group at the Space Research Institute in Moscow, and he is a professor at Moscow State University. He is the author or coauthor of numerous fundamental articles and books in many branches of physics and applied physics.

I. D. Novikov was a key member of Zel'dovich's research group from 1963 until 1979. He now leads his own astrophysics research group at the Space Research Institute in Moscow and is a professor at the Moscow Physical Technical Institute. He has published numerous fundamental articles on relativity theory, astrophysics, and cosmology.

Gary Steigman is professor of physics at the Bartol Research Foundation of the Franklin Institute, University of Delaware.

Relativistic Astrophysics, Volume 2, has been translated from the original Russian *Relyativistskaya Astrofizika: Stroeniye i Evolyutsiya Vselennoi* (Moscow: Izdatel'stvo "Nauka," 1975) as revised and enlarged by the authors.

The University of Chicago Press, Chicago 60637
The University of Chicago Press, Ltd., London

© 1983 Ya. B. Zel'dovich and I. D. Novikov
© 1983 by The University of Chicago
All rights reserved
Published 1983
Printed in the United States of America
90 89 88 87 86 85 84 83 5 4 3 2 1

Library of Congress Cataloging in Publication Data

Zel'dovich, ÎAkov Borisovich.
Relativistic astrophysics.

Translation of Reliativistskaià astrofizika.
Vol. 2: Translated by Leslie Fishbone; edited by Gary Steigman.
Includes bibliographies and indexes.
CONTENTS: v. 1. Stars and relativity.—v. 2. The structure and evolution of the universe.
1. Astrophysics. 2. Relativity (Physics). 3. Cosmology. 4. Gravitation. I. Novikov, Igor' Dmitrievich. II. Title.
QB461.Z4413 523.01 77-128549

ISBN 0-226-97955-5 (v. 1) AACR1
ISBN 0-226-97957-1 (v. 2)

CONTENTS

Editor's Foreword, xiii

Preface to the English Edition, xv

Introduction, xix

I. THE HOMOGENEOUS, ISOTROPIC UNIVERSE: ITS EXPANSION AND GEOMETRICAL STRUCTURE

1. LOCAL PROPERTIES OF THE HOMOGENEOUS, ISOTROPIC COSMOLOGICAL MODEL, 3

1.1 The Local Law for the Velocity Distribution, 3
1.2 The Law of Evolution and the Critical Density, 6
1.3 The Duration of the Expansion, 9
1.4 Two Particular Solutions; the Early Stage, 9
1.5 The Effect of Pressure on the Expansion Law: Qualitative Considerations, 12
1.6 The Equation of Motion, 14
1.7 The Expansion Time, 15
1.8 The Early Stages, 17
Appendix to § 1.3, 18

2. RELATIVISTIC THEORY OF THE HOMOGENEOUS, ISOTROPIC UNIVERSE, 20

2.1 Einstein's Gravitational Equations and Friedmann's Cosmological Equations, 20
2.2 Global Geometrical Structure of a Model of the Universe: The Space of Constant Positive Curvature, 23
2.3 The Metric of the Open Model, 28
2.4 The Limiting Case of Small Matter Density, 29
2.5 The Critical-Density Case, 31

3. THE PROPAGATION OF PHOTONS AND NEUTRINOS; OBSERVATIONAL METHODS FOR TESTING COSMOLOGICAL THEORIES, 33

3.1 The Redshift and Momentum Decrease, 33
3.2 Observable Quantities and the Horizon, 35
3.3 Graphs and Formulae for Functions Which Determine Observable Quantities, 41

3.4 Working Formulae as Functions of the Redshift z, 52
3.5 The First Approximation and Euclidean Space; the Deceleration Parameter, 56
3.6 The Distribution with Respect to Apparent Magnitude, 61
3.7 On the Possibility of Choosing among Cosmological Models by Observations of Distant Objects, 66
3.8 The Evolution of Radio Sources, 70
3.9 The Determination of H_0 and q_0 from Observations, 78
3.10 Observable Quantities in a Universe That Is Homogeneous Only on the Average, 85
3.11 The Kinetic Equations for Photons, 89
3.12 Is the Redshift Unambiguously Explained by the Expansion of the Universe? 96

4. THE COSMOLOGICAL CONSTANT, 99
4.1 Does the Cosmological Constant Differ from Zero? 99
4.2 Cosmological Models with the Λ Term, 102
Appendix to Part I: The Semiclosed Model and the Physical Interpretation of the Kruskal Metric, 109

II. PHYSICAL PROCESSES IN THE HOT UNIVERSE

5. INTRODUCTION TO PART II, 117
5.1 Introductory Remarks and Historical Review, 117
5.2 Electromagnetic Radiation in the Universe—a Review of the Observations, 125
5.3 Observational Proof of the Existence of the Relic Radiation (RR), 129

6. THERMODYNAMIC EQUILIBRIUM AT THE BEGINNING OF THE COSMOLOGICAL EXPANSION, 134
6.1 Basic Epochs in the Evolution of the Hot Universe, 134
6.2 The Cosmological Expansion of the High-Temperature Plasma and the Condition of Thermodynamic Equilibrium, 136
6.3 The Hadron Era in the Evolution of the Universe, 139
6.4 Hagedorn's Theory, 147
6.5 The Equilibrium Concentration of Nucleons and Antinucleons in the Charge-asymmetric Universe, 151

7. THE KINETICS OF ELEMENTARY PARTICLE PROCESSES, 153
7.1 Neutrinos in the Theory of the Hot Universe, 153
7.2 Cosmological Gravitational Waves, 157
7.3 Antinucleons in the Hot Model, 163
7.4 Relic Quarks in the Hot Model, 170

7.5 Nucleosynthesis in the Hot Model, 173
7.6 A Comparison of the Observational Data on the Distribution of Light Elements in the Universe with the Predictions of Theory, 183

8. THE RADIATION-DOMINATED PLASMA AND THE RELIC RADIATION, 189

8.1 Introduction and General Review, 189
8.2 Recombination: Equilibrium and Kinetics, 195
8.3 The Interaction between the Electrons and the Photons in the Dilute Plasma, 200
8.4 The Effect of the Electrons on the Radiation Spectrum, 207
8.5 Early Energy Release and Quasi-Equilibrium, 209
8.6 When Does the Relic Radiation Arise? 215
8.7 Late Energy Release, 215
8.8 The Intergalactic Gas: Its Radiation and Density, 220
8.9 Radio Emission by the Ionized Intergalactic Gas and the Epoch of Neutral Hydrogen, 226
8.10 The Interaction of Cosmic Rays with the Relic Radiation, 229

III. GRAVITATIONAL INSTABILITY IN COSMOLOGY; GALAXY FORMATION

9. INTRODUCTION TO PART III, 237

10. GRAVITATIONAL INSTABILITY IN NEWTONIAN THEORY, 240

10.1 Jeans's Theory, 240
10.2 The Instability of Expanding, Homogeneous Matter, 247
10.3 Large-Scale Perturbations; the Self-similar Solution, 251
10.4 A View of Perturbations as Variations in the Parameters of a Solution; the Evolution of Perturbations in a Nonflat ($\Omega_0 \neq 1$) Model, 253
10.5 Formulae Describing the Evolution of Perturbations, 255
10.6 Numerical Estimates, 259
10.7 The Instability of a Collisionless, Self-gravitating Gas, 262

11. INSTABILITY IN THE HOT MODEL, 265

11.1 Conditions for the Growth of Perturbations, 265
11.2 Dissipative Processes and the Decay of Adiabatic Perturbations, 270
11.3 The Interaction of Perturbations with Free Particles, 274
11.4 Entropy Perturbations, 276
11.5 Rotational Perturbations, 278
11.6 Matching Perturbations When the Equation of State Changes, 280

12. GRAVITATIONAL INSTABILITY IN THE GENERAL THEORY OF RELATIVITY, 285

12.1 General Principles and Equations, 285
12.2 Classification of the Perturbations, 289

12.3 Scalar Perturbations, 291
12.4 Vector (Rotational) Perturbations, 295
12.5 Tensor Perturbations and Gravitational Waves, 299
12.6 Entropy Perturbations in Relativistic Theory, 302
12.7 The Quasi-isotropic Solution and the Hypothesis of a Uniform Distribution of Perturbations, 303
12.8 Long-Wavelength Perturbations and Their Representation in Terms of Spherical Waves, 305

13. STATISTICAL THEORY, 310
13.1 Randomness and Fourier Analysis, 310
13.2 The Correlation Function and the Sizes of Self-gravitating Objects, 311
13.3 Deviations from the Average Density in a Given Volume, 317
13.4 The Limitations and Complexity of the Linear Theory, 322

14. THE NONLINEAR THEORY OF PERTURBATIONS AND THERMAL INSTABILITY, 325
14.1 Perturbations in a Matter-Dominated Model; Problems Admitting Exact Solutions, 325
14.2 Perturbations in a Matter-Dominated Model; an Approximate Analysis of a General Asymmetric Case (the "Pancake" Model), 329
14.3 Nonlinear Spectral Theory, 335
14.4 The Emergence of Long-Wavelength Perturbations from a Gas of Stars or Star Clusters, 339
14.5 Thermal Instability and the Separation of a Homogeneous Gas into Phases, 344
Appendix to Section 14.2, 347

15. THEORIES OF GALAXY FORMATION, 350
15.1 Introduction, 350
15.2 Adiabatic Perturbations: Prerequisites, 358
15.3 Shock Waves, 360
15.4 Thermal Processes in the Compressed Gas, 366
15.5 Cluster Masses and Protocluster Fragmentation, 372
15.6 Galaxy Rotation, 377
15.7 Galactic Magnetic Fields, 384
15.8 The Theory of Entropy Perturbations, 389
15.9 The Vortex Theory, 394
15.10 A Comparison of the Evolutionary Theories of Galaxy Formation, 399
15.11 Observational Data concerning the Properties of Galaxies and Galaxy Clusters; the Average Density of Matter in the Universe, 405

15.12 Astrophysical Consequences of the Existence of a Heavy Neutral Lepton, 413
Mathematical Appendix to § 15.9, 416

16. RESEARCH ON PERTURBATIONS USING THE RELIC RADIATION, 419

16.1 Introduction, 419
16.2 Antimatter Annihilation, 422
16.3 Adiabatic Perturbations, Acoustic Oscillations, and Their Effect on the Relic Radiation Spectrum, 428
16.4 Perturbations of the Spatial Homogeneity and Isotropy of the Relic Radiation, 432
16.5 The Detection of Density Perturbations Using the Relic Radiation, 438
16.6 The Perturbation Spectrum and the Low-Density, Hyperbolic Model, 443
16.7 The Angular Distribution of the Relic-Radiation Fluctuations, 444

17. GRAVITATIONAL WAVES IN COSMOLOGY, 449

17.1 Introduction, 449
17.2 General Information about Gravitational Waves, 450
17.3 Gravitational Waves in the Theory of Small Perturbations of a Cosmological Model, 454
17.4 The Expected Intensity of Relic, Short-Wavelength Gravitational Radiation, 457
17.5 The Equipartition Theorem and Long-Wavelength Gravitational Radiation, 460
17.6 The Generation of Gravitational Waves in the Present Epoch, 461
17.7 The Effect of Gravitational Waves on the Relic Radiation, 465
17.8 The Peculiar Motion Induced by Gravitational Waves, 467
17.9 The Interconversion of Gravitational and Electromagnetic Waves, 470

IV. ANISOTROPIC COSMOLOGY

18. INTRODUCTION TO PART IV, 477

19. VERY SIMPLE ANISOTROPIC COSMOLOGICAL MODELS, 480

19.1 The Newtonian Theory of a Very Simple Homogeneous, Anisotropic (HA) Model as the Limiting Case of a Local Problem, 480
19.2 A Paradox Involving Newtonian Gravitational Theory, 484
19.3 A Simple Relativistic Model: the "Vacuum" Solution near a Singularity, 485
19.4 A Comparison of the Newtonian and Relativistic Problems, 487

20. MATTER IN AN ANISOTROPIC COSMOLOGICAL MODEL, 490

20.1 Isotropization of a Model with a Stress-Energy Tensor of the Pascal Form, 490
20.2 The Effect of a Spatially Anisotropic Stress-Energy Tensor on a Cosmological Model, 491
20.3 Cosmological Models with a Homogeneous Magnetic Field, 495
20.4 Perturbations in a Homogeneous, Anisotropic (HA) Universe, 501
20.5 The Instability of the Cosmological Models with Respect to the Emergence of Overall Matter Motions, 506

21. THE PHYSICS OF PROCESSES DURING THE EARLY STAGES OF EXPANSION IN ANISOTROPIC MODELS, 511

21.1 Weakly Interacting Particles in an Anisotropic Cosmological Model, 511
21.2 Neutrinos in an Anisotropic Model, 516
21.3 The Effect of Viscosity on the Expansion of Anisotropic Models, 519
21.4 The Kinetic Theory of Neutrinos in an Anisotropic Model; the Self-similar Solution, 521
21.5 Primordial Nucleosynthesis in Anisotropic Models, 523

22. A GENERAL ANALYSIS OF HOMOGENEOUS COSMOLOGICAL MODELS, 530

22.1 The Concept of the Homogeneity of a Cosmological Model, 530
22.2 A Differential Criterion for Homogeneity, 535
22.3 The Dynamical Properties of Homogeneous Models near a Singularity, 537
22.4 The "Mixmaster" Model, 540
22.5 On the Impossibility of Mixing in the Mixmaster Model, 545
22.6 Quantum Limitations on the Mixmaster Model, 547
22.7 The Isotropization of Homogeneous Cosmological Models in the Course of Expansion, 548
22.8 The Anisotropy of the Relic Radiation (RR) in Bianchi Type I Models with the Critical Matter Density, 555
22.9 The Expected Anisotropy of the Relic Radiation (RR) in Homogeneous, Anisotropic (HA) Models with a Curved Three-Dimensional Space, 559

V. THE SINGULARITY AND NEW THEORETICAL DEVELOPMENTS

23. THE COSMOLOGICAL SINGULARITY, 569

23.1 Introduction, 569

23.2 The Singularity at the Beginning of the Expansion, 571
23.3 The General Cosmological Solution with a Singularity, 577

24. PHYSICAL PROCESSES NEAR A SINGULARITY AND DEVELOPMENTS IN THE THEORY OF GRAVITATION, 581

24.1 Introduction, 581
24.2 Cosmological Consequences of Hagedorn's Theory, 583
24.3 Quantum Phenomena That Take Place near Singular States of a Metric and in Strong Gravitational Fields, 585
24.4 The Creation of Charged Particles in Electrodynamics, 594
24.5 The Mathematical Theory of Particle Creation, 603
24.6 Superspace and Mini-Superspace, 610
24.7 The Hypothesis of Baryon Nonconservation and the Charge Asymmetry of the Elementary Particles, 613
24.8 The Cold Universe and the Perturbation Spectrum, 616
24.9 The Steady-State Theory of the Universe, 622
24.10 Mach's Principle and the Coincidences among Large Numbers Occurring in Physics and Cosmology, 625
24.11 The General Theory of Relativity (GTR) and the Topological Structure of the Universe, 633
24.12 Local Topology, White Holes, Black Holes, and Cosmology, 641
24.13 Statistical Physics and Gravitation, 645
24.14 The Brans-Dicke Theory of Gravitation (BDT) and Its Cosmological Consequences, 648
24.15 Cosmology and New Hypotheses in Field Theory, 654
24.16 An Oscillating Universe? 658
24.17 The Creation of Gravitons near the Singularity, 663
24.18 The Singularity and Conformal Invariance, 666
24.19 The Direction of Time, 670

REFERENCES, 675

AUTHOR INDEX, 699

SUBJECT INDEX, 705

EDITOR'S FOREWORD

In less than two decades, modern cosmology has emerged as one of the most exciting, vibrant, and productive fields of scientific endeavor. As midwives in its birth and as shepherds of its more recent growth to maturity, Ya. B. Zel'dovich and I. D. Novikov have played a role of incalculable importance. It is no hyperbole to say that they have conceived of, and carried through, most of the developments which have led to "modern" or "physical" cosmology. This book, the second volume of *Relativistic Astrophysics*, has been long awaited. The wait has been worth it.

In its approach to modern cosmology, this book is encyclopedic. I can think of no important developments which are omitted. It should be an invaluable addition to every physicist's bookshelf. It will serve well those who wish to learn as well as those who wish to review. The approach is a pedagogic one in which no details are omitted and the physics (rather than the mathematical formalism) is always at the forefront.

The publication of this eagerly anticipated book has been much delayed. Virtually all the blame in this respect must fall on me as editor; some fault, however, belongs to the authors. I began my task in 1972 with quite a different manuscript. The subject of cosmology was then in rapid growth, and Zel'dovich and Novikov would provide new chapters as rapidly as I edited the old ones. The process was not converging. A halt was called, and Zel'dovich and Novikov produced an entirely new book, *Cosmology* (Nauka), in Russian which had to be translated before the editing could begin anew.

As editor, I have made a real attempt to verify the correctness of the text and to ensure its clarity. This, most assuredly, is the major cause of the delay in publication. Although I hope the time has been well spent, I urge the serious reader to approach the text with care; there may still be minor errors which have escaped the diligence of the authors and the editor.

As editor, I am indebted to Dr. Leslie Fishbone who provided a superior translation, greatly simplifying my task. Dr. Fishbone worked on the manuscript while at the Department of Physics and Astronomy of the University of Maryland and at the Department of Physics of the University of Utah. He wishes to express his appreciation to the faculty and staff of both departments for their support and assistance. He particularly wishes to thank Anita Wellin for translating the references.

Phyllis Jackson typed most of the manuscript, and I am grateful to her for her patience and her competence. Thanks too are due Bernard Carr, Doug Macdonald, Ian Redmount, and Mike Wilson for editorial assistance. Finally, I must thank Kip S. Thorne for always being available to help when I needed it.

GARY STEIGMAN

PREFACE TO THE ENGLISH EDITION

This book constitutes the second volume of our monograph *Relativistic Astrophysics.* It is devoted to cosmology and can be read independently of the first volume, *Stars and Relativity* (Zel'dovich and Novikov 1971).

Modern cosmology is a vast, rapidly developing field of knowledge. Its theoretical basis is the set of cosmological models devised by the Soviet mathematician Alexander Friedmann. Its observational basis is the work that led to the discovery by the American astrophysicist Edwin Hubble of the law describing the redshift in the spectra of galaxies.

Though the kinematics of the evolving Universe became known decades ago, only in more recent years did research into the physics of processes occurring in the expanding Universe receive a reliable observational and theoretical basis. Indeed, confirmation of the so-called hot-Universe theory has come only during the past decade. Now, with respect to its overall features, one can consider as known the present state of the Universe and that of its most recent past. The results achieved are of permanent significance, forever to be counted in the golden stock of science.

However, these achievements have led to the emergence of new problems, on which an unusually active assault has begun. Diverse attempts have been undertaken to work out the deepest and most difficult questions, those relating to the singular state in the very distant past of the Universe, to the origin of galaxies, and to other matters.

Research concerned with the cosmological singularity is of particular theoretical significance because, though the singular state is distant from us by a time greater than 10^{10} years, every particle (or its predecessor) has emerged from the crucible of the singularity. Said otherwise, the present and future states of the Universe depend upon its past, and consequently on the singular state.

Modern cosmology makes use of the entire arsenal of knowledge in both physics and astronomy. The aforementioned classical hot-Universe theory utilizes the general theory of relativity, thermodynamics, hydrodynamics, and plasma theory. In studies of the singularity, wide use is made of such younger branches of physics as quantum-field theory. It is therefore natural to find as neighbors in this book respected and established theories more than 50 years old and hypotheses which have hardly emerged from infancy. Sometimes they contradict each other!

The parts of this monograph in which are described the basic observational

facts and classical theories should form part of the general-education background of every astronomer—and even, in our opinion, of every physicist. On the other hand, knowledge of the entire content of the book is necessary only for those who are working or intending to work in the field of cosmology. Though not a substitute for the original literature, this monograph gives a sufficiently complete picture of the present state of cosmology, especially theoretical cosmology, to satisfy most.

Combining the heterogeneous problems of cosmology in a single volume seemed to us unusually difficult. What seemed utterly unrealistic was our original intention to redo for this edition the cosmology section of our Russian-language book *Relyativistskaya Astrofizika* (Zel'dovich and Novikov 1967*b*). A monograph had to be written all over again. In spite of the large size of the resulting book, it was only with great difficulty that we succeeded in covering all parts of modern cosmology. As in our previous books, along with the mathematical theory of processes, we have tended to give a clear interpretation of the theory, showing especially how the formalism of the theory is applied to the solution of concrete problems.

The Russian version of this book was published in 1975. The volume in hand represents a translation of the 1975 Russian edition. In the course of the work on the translation, however, we once again had to redo several sections in order to describe the achievements of the last few years.

To conclude our excuses regarding the difficulties involved in writing a monograph of this sort, we present the following topical citation from the introduction to Kepler's book *The New Astronomy Based upon Causes, or The Physics of the Heavens:*

"At the present time, the gravest fate is reserved for those who write mathematics or, especially, astronomy books. If the necessary strictness in terminology, explanations, proofs, and conclusions is not maintained, then the book will not be mathematical. If this strictness is maintained, then it will be very tiring to read the book, especially in Latin, which is devoid of the charm peculiar to text written in Greek. Therefore, it is now very rare that one encounters suitable readers; the majority prefers to avoid reading altogether."

While working on this monograph, we enjoyed the unfailing support of our comrades in the theoretical astrophysics section of the Institute of Applied Mathematics of the Academy of Sciences of the U.S.S.R. and of the Institute for Cosmic Research of the Academy of Sciences of the U.S.S.R. We mention particularly A. G. Doroshkevich and R. A. Sunyaev, together with whom Chapter 8 was written. Highly useful too were discussions and debates with workers at the Institute for Cosmic Research, the Shternberg State Astronomical Institute, the Physical Institute of the Academy of Sciences, the Institute for Theoretical Physics, the Institute for Physical Problems, and other institutes, and also with foreign colleagues.

We want to thank especially our foreign colleagues Dr. Leslie Fishbone,

the translator, and Dr. Gary Steigman, the editor of the translation, for having made our book accessible to English-language readers. Dr. Steigman made particularly important contributions by commenting on the substance of the text and by correcting several misprints and errors in the Russian edition. Finally, we hope that the joint work by the editor and us on the translation itself, both through correspondence and during Dr. Steigman's visit to us in Moscow, has improved the English edition.

YA. B. ZEL'DOVICH
I. D. NOVIKOV

INTRODUCTION

Scientific cosmology arose significantly earlier in history than did other astronomical disciplines. In spite of this earlier flowering, and indeed, perhaps as a consequence of it, the subject is presently being developed particularly intensively. Discoveries of enormous significance have resulted.

In the following short review, we shall try to display the structure and logic of modern cosmology. In so doing, it is not our intention to give a historical account. As a result we shall refrain from giving references to the literature; these will be presented when a detailed account of the question at issue is presented. Our philosophy is that the history of the Universe is infinitely more interesting than the history of the study of the Universe.

The most important fact of cosmology is the observed isotropy of the Universe. In other words, the visible picture of the Universe is independent of the direction in which an observer looks. This applies to the radio emission penetrating the entire Universe (the "relic radiation") to which we shall return for a more detailed discussion; to the longer wavelength radio emission from discrete sources, when averaged with respect to an area on the sky containing many sources; and to the X-ray emission, the origin of which is not entirely clear. These forms of radiation reach us from great distances, yet the deviation from isotropy in their intensity does not exceed 0.1–1.0%.

The next step in understanding the structure of the Universe consists of a bold extrapolation. We know that we are located in a spiral galaxy, which in turn is part of a cluster of galaxies. We see other clusters of galaxies scattered in space. However, the isotropy of the relic radiation (RR) and of the other forms of radiation incident from afar is more fundamental than the local inhomogeneity of our immediate surroundings. Indeed, the distribution on the sky of galaxy clusters also presents on the average an isotropic picture. It is therefore natural to assume that the Universe is homogeneous on large scales, although inhomogeneous on small scales. The observed isotropy of the radiation gives a quantitative estimate of the degree of homogeneity. The deviations from homogeneity turn out to be less than 0.1–1.0% on a scale of $\sim 10^{10}$ light-years.

It is necessary to emphasize that, at first glance, the observational data do not by themselves contradict an assumption of spherical symmetry for the Universe. It is possible to assume that the Earth (the solar system or the Galaxy) is located at the center and that the average density of the matter

in the Universe depends on the distance from the center, but not on the direction.

The selection of a homogeneous model instead of a model with a center in the Galaxy is based on general philosophical reasoning: In the Middle Ages it seemed natural simply to assume that the Earth is in the central position. Since that time, the prevailing philosophical views have changed; we have become more modest. Finally, an in-depth analysis of all of the observational data leads to the conclusion that the Universe is homogeneous.

Accepting the homogeneous model of the Universe, we are led to several conclusions which follow from well-known physical laws. But it is also appropriate to ask: How justifiable is the use of known physical laws established in the laboratory? Should one expect, upon passing to the grand scale of the Universe, that these laws must be changed? Such a change in the laws of physics over large scales did occur when the general theory of relativity (GTR), with its notion of spacetime curvature, was developed. Curvature clearly must be manifested over large cosmological scales. Now GTR is so designed that the correspondence principle is satisfied: it gives the known laws of physics identically over small scales. Of course, these known laws have been verified with great accuracy over laboratory and even solar-system scales. However, even within the limits of the solar system (and in the most refined, accurate experiments in the laboratory), GTR predicts specific corrections whose existence has been experimentally confirmed. Finally, besides the demonstrated existence of spacetime curvature, there is still the unanswered question that applies specifically to large scales—the question of the cosmological constant.

The cosmological constant describes the (positive or negative) energy density and corresponding pressure attributed to the vacuum. This leads to the appearance of an additional relative acceleration (positive or negative) between any two particles, an acceleration depending only on the distance between the particles and not on their mass. This is a very small addition, however. The cosmological constant describes a universal force of repulsion or attraction proportional to distance. This change in the law of gravitation cannot be measured in laboratory experiments. However, for the dynamics of the Universe as a whole, because of the enormous distances involved, the effects of the cosmological constant can be substantial.

The appearance in physics of new laws and new formulations has always been connected with old laws falling into irreconcilable contradiction with experiment or proving to be logically inconsistent when applied to a new region of experience. So it was, for example, when prequantum physics collided with the ultraviolet catastrophe in the radiation law and with the unexplained stability of atoms. So it was, also, when a relativistic generalization of Schrödinger's equation was required for a description of high-velocity quantum phenomena. Let us now approach GTR in this spirit. Applying GTR to the unbounded Universe, we confront neither logical

contradictions in the theory itself nor any glaring contradictions between theory and observations. In particular, there are no observational data suggesting a limitation on the applicability of GTR to the scales of the Universe.[1] Therefore, the assumption that a change in GTR is needed in applications to cosmology is unfounded.

Thus, the aggregate of theoretical, experimental, and observational facts stand in favor of the applicability of the physical laws and GTR to a description of the evolution of the Universe from *almost* the very beginning of the expansion. They apply from times when the matter density is much greater than the density of nuclear matter, ($\rho > 10^{14}\,\mathrm{g\,cm^{-3}}$), up to the present time.

The authors want to emphasize the "orthogonality" of their point of view from that of some others. We do not agree with the "theories" appearing from time to time with features that violate the fundamental laws of physics. Such theories are, for example, those in which there is constant creation of matter "out of nothing" far from the singularity (the theory of the steady-state universe), or those which involve a decrease of the gravitational constant with time.[2] Concrete reasons for our disagreement and a criticism of such theories are given in Part V of the book. In short, we adopt the viewpoint that the homogeneous and isotropic Universe can be examined within the realm of GTR.

As is well known, there exists an extensive family of solutions to the Einstein equations for the evolution of the Universe which preserves the homogeneous and isotropic properties in the course of time. But what, specifically, are these properties?

Observations show that we live in an expanding, indeed an evolving, Universe. The redshift of the spectral lines of light from distant galaxies is a consequence of the recession from us of these galaxies. The latest measurements confirm the proportionality between the velocity and the distance: $v = H_0 r$, where $H_0 \approx 50\,\mathrm{km\,s^{-1}\,Mpc^{-1}} = (6 \times 10^{17}\,\mathrm{s})^{-1} = (2 \times 10^{10}\,\mathrm{yr})^{-1}$. The quantity with the dimensions of time inside the parentheses corresponds roughly to the time for expansion from the high-density state. The quantity H_0—"Hubble's constant" (the present value of the "Hubble parameter")—is known from observations to an accuracy of at least 50%.

Many solutions of the equations of GTR agree with the linear relation $v = H_0 r$. Indeed, such an expansion law is necessary in order that the most general properties of such a solution—homogeneity and isotropy—be preserved for the duration of the evolution. At the same time one also assumes

1. As we shall see, relativistic cosmology leads to the necessity for a state in the past with enormous density and with enormous spacetime curvature. This is the so-called singular state. In such a situation, GTR in its existing form is not applicable. A more detailed discussion of this will appear later.

2. We do not exclude unexpected properties of matter at high temperature and density, subject to energetic and thermodynamic limitations.

that the density and pressure of the matter are everywhere the same. That is, they do not depend on the spatial coordinates, though they do depend on time. For a complete solution to the problem of specifying a solution of the equations, however, it is not sufficient to give these properties and the value of H_0 alone. It is also necessary to know the numerical values of the matter density and the pressure today, and the value of the cosmological constant Λ.

Very likely, $\Lambda \approx 0$. In this case one can show that for any matter density, the duration of the past evolution is less than H_0^{-1}. The future evolution then depends essentially on ρ_0, the present-day value of the matter density only. There is a critical value of the density, ρ_c, depending on H_0 (for $H_0 = 50$ km s^{-1} Mpc^{-1}, this critical value is $\rho_c \approx 0.5 \times 10^{-29}$ g cm^{-3}) such that for $\rho_0 < \rho_c$ the expansion will continue without bound, while for $\rho_0 > \rho_c$ the expansion will change to contraction.

The most important property of the Universe—its geometrical structure, whether it is finite or infinite—also depends on the density alone (if $\Lambda \approx 0$). For $\rho_0 > \rho_c$, the Universe is finite although unbounded. This is similar, for example, to the surface of a sphere, which is finite but has no boundaries. For $\rho_0 < \rho_c$, the Universe is infinite and, in this sense, does not differ from classical, three-dimensional, Euclidean space. Thus the density, the future, and the geometrical structure of the Universe are related to one another.

The case of a nonzero cosmological constant is more complicated. For example, given a definite relation between Λ and ρ_0, the finite-Universe case is possible, though in the future, unlimited expansion is in prospect. Here, in a short review, it is inappropriate to present a complete classification of such solutions.

As we shall see later, the direct determination of the average density in the Universe from observations of diverse objects is a very difficult problem since many forms of matter are difficult to observe. This is, however, a major aspect of cosmological research.

In principle, one can determine the structure of the Universe from the curvature of space by observing how the brightness of sources with known absolute luminosity (e.g., galaxies) changes in relation to their distance or to their redshift. One can use the redshift because it is directly connected to distance. However, this redshift-distance dependence varies from model to model.

The theory of light propagation and the connections between the properties of distant objects and observable quantities are those aspects of cosmology which have been worked out most extensively. It would, in principle, be sufficient to measure the redshift and the light flux of two distant objects with known absolute luminosity in order to determine H_0 and the density ρ_0. This is true only when $\Lambda = 0$; however, if $\Lambda \neq 0$, an additional observation is necessary. In order to establish the structure of the Universe, one can measure in place of the luminosity the number of objects of a given type

in a given element of solid angle and in a specified interval of redshift. Again though, this must be for objects whose density in space is known from independent considerations. However, this simple and clear program has to this day not been carried out—and not only because of practical, technical difficulties. A deep problem of principle arises because celestial bodies evolve. In observing distant objects, we are also observing them in the distant past. Even for objects of known type (galaxies and quasars), we therefore cannot know with certainty their absolute luminosity and number per unit volume at the instant when the radiation, received by us today, was emitted.

Therefore, a comparison of the theory with observations has not led to the establishment of the structure of the Universe. Rather, proof has been obtained that the specific number density[3] (or the intrinsic luminosity) of the most powerful sources of energy in the Universe—quasars and radio sources—was substantially higher in the past than it is today. For ordinary galaxies, one expects much greater stability and smaller evolutionary effects. But the distances over which galaxies can be observed are small compared to cosmological scales. To this day, therefore, the structure of the Universe has not been established through observations of ordinary galaxies either.

For a determination of the structure of the Universe, it thus became necessary to return to the difficult problem of finding the average density by means of direct observations of the various celestial bodies and of the radiation in the Universe. Indeed, these observations are necessary for one additional reason.

The development of cosmology is inseparable from the question of the material content of the Universe. For the dynamics of the Universe, it is sufficient to know only one quantity, the average density. But, for an understanding of the physical processes, it is necessary to know which particles fill the Universe. We are interested in both questions.

The average density of visible matter in galaxies today is known approximately. If one "spreads" it over all of space, then the resulting density is about 10^{-31} g cm^{-3}, i.e., about 50 times less than the critical density. Very recently it has been shown that massive elliptical galaxies are possibly surrounded by halos consisting of stars of low luminosity. Taking account of this, the average density can be increased several-fold. Now the mass of galaxies is determined from an analysis of the motions of stars and gas clouds under the natural assumption that the age of the galaxies is many times larger than the period of a star in a galactic orbit. Such a method permits one to determine the total mass of a galaxy, a mass including that of invisible forms of material such as extinct stars.[4] Since the mass of a galaxy

3. That is, the number of sources taken over a fixed mass of matter in the Universe.

4. Notice, however, that this applies to the mass within the galaxy. Mass in the hypothetical halo surrounding the galaxy does not affect motion in the inner part.

is approximately equal to the sum of the rest masses of the baryons (neutrons and protons) of which the galaxy is composed, one can thereby obtain the average number density of baryons in the Universe today—a density whose order of magnitude is $\sim 6 \times 10^{-8}$ cm^{-3}.

One common assumption concerns the charge symmetry of the Universe, i.e., the existence of equal amounts of matter and antimatter on the average. It is necessary to emphasize that, from the point of view of the theory of elementary particles, such an assumption is *a priori* possible but certainly not required. Special searches have been undertaken for the effects of the annihilation of particles and antiparticles between matter galaxies and antimatter galaxies. These searches have yielded a negative result. Therefore, at the present time the assumption of charge asymmetry in the Universe is more probable. This assumption means, for example, that there is only an insignificant number of antibaryons among the cosmic rays.

Observations show the presence of intense radio emission in the centimeter and millimeter bands which cannot be ascribed to discrete sources. It has been assumed, therefore, that this radiation is primordial and arises in the very early stages of the Universe.

Using as a basis the observational data and a minor theoretical extrapolation, the number density of the primordial (relic) photons has been calculated to be about 400 cm^{-3}, which is 10^8–10^{10} times greater than the average number density of baryons in the Universe. This RR corresponds to blackbody radiation at a temperature of 2.7 K. The average energy of the photons is about 0.0007 eV, while the energy density of the radiation is $\epsilon = 4 \times 10^{-13}$ ergs cm^{-3}. The corresponding mass density is $\rho_{\text{rad}} = \epsilon/c^2 = 5 \times 10^{-34}$ g cm^{-3}. Thus the contribution of the radiation to the total density is quite small at the present time.

The numerous theoretical and observational studies of the intergalactic gas are still incomplete. The neutral hydrogen content of this gas is quite low, for any such gas must be almost entirely ionized. Its temperature presumably lies within the limits 2×10^4 K $< T < 3 \times 10^8$ K. Such gas is practically transparent at all wavelengths; indeed, the gas itself would emit somewhere between the ultraviolet and X-ray part of the spectrum. The best estimates give the inequality $\rho_{\text{gas}} \lesssim \rho_c$.

A direct determination of the density of neutrinos and gravitational waves is quite difficult. Their density could exceed by many times the density of ordinary matter (baryons). Even so, direct physical methods would still not have sufficient sensitivity to detect these neutrinos and gravitons. Indirect estimates suggest that the density of such particles is substantially less than the density of ordinary matter.

At present, there are no reliable estimates of the possible numbers of extinct quasars, stars, and other low-luminosity or nonluminous objects between galaxies. Concerning galaxies though,the average density of matter in the Universe contributed by them has been fairly reliably established.

The value is $\rho_{gal} \approx 10^{-31}$ g cm$^{-3} \approx 0.02\ \rho_c$. Concerning other forms of matter, one can only say that well-studied forms (for example, RR) have a density substantially less than ρ_c. The density of the remaining forms is very poorly known.

Thus, there is as yet no answer to the question of whether or not ρ_0 is larger than ρ_c, and, consequently, whether or not the Universe is finite or infinite.

Knowing the present content of the Universe, one can trace the earlier stages of evolution. The most important fact is that of the expansion of the Universe. As the expansion proceeds, the RR temperature falls. The value of this temperature today is 2.7 K; in the past it was much higher. Early in the expansion, the dense matter is opaque to radiation. The matter and radiation are in thermodynamic equilibrium and have a very high temperature. Hence the name, the "theory of the hot Universe."

One can consider as firmly established the general picture of evolution that is encompassed by the name, the "theory of the hot Universe." This picture entails first of all the homogeneous (the same at all points in space) and isotropic (the same in all directions) expansion of the Universe. Second, it entails a Universe filled with material that is numerically dominated by photons. Thus, knowing the expansion law and making use of the laws of physics, one can compute the past history of the matter and determine the ongoing physical processes.

The mass density of ordinary matter (baryons) falls with the expansion inversely as the volume, $\rho_b \sim V^{-1}$. The RR energy density falls faster than ρ_b: $\rho_{rad} \sim V^{-4/3}$. Consequently, in the past, the photons predominate at an early stage, not only with respect to number but also with respect to mass. At this stage, called the radiation-dominated (RD) era, the ordinary matter consists of completely ionized helium and hydrogen.

From the equations of mechanics and from the known interactions between photons and atoms, one can find the temperature and composition as a function of time (where we will reckon the time t from the moment when $\rho \to \infty$ in the Friedmann solution). For example, at the time $t = 1$ s, the temperature is about 1 megavolt or 10^{10} K, while the density is 10^6 g cm^{-3}. Besides photons, there are at this time almost as many pairs of electrons and positrons present. However, complex nuclei cannot exist at this time. Protons and neutrons exist in almost equal numbers due to the collisions with electrons and positrons, collisions which lead to mutual transformations of the protons and neutrons into each other.

As the expansion proceeds, the positrons disappear. A portion of the neutrons decay, while the remaining neutrons combine with the protons, yielding finally 70% hydrogen and 30% helium by mass. Traces of deuterium and of helium-3 may result, but there is almost a complete absence of heavier elements. This prediction of the theory does not contradict the meager evidence on the possible content of primordial matter (matter not

processed by stellar nucleosynthesis). Neutrinos and antineutrinos from this period must also remain. Moreover, those remaining must be roughly equal in number to the number of photons and must have an average energy today corresponding to a temperature of several degrees.

During the expansion following the RD era, i.e., in an epoch close to ours, the mass density of ordinary matter exceeds the mass density of photons. Moreover, this matter is in the form of neutral atoms.

In this general picture many details are lacking. These include a treatment of deviations from homogeneity and a treatment of the earliest stages, when, along with the photons, diverse particles and antiparticles are present and quantum phenomena are important.[5] As we shall see later, there is a definite connection between the question of inhomogeneity and that of the evolution during the earliest stages.

From the observational point of view, the possibility of studying the early stages is limited by the scattering of electromagnetic waves by electrons. All RR received today is repeatedly scattered by electrons when the scale of the Universe is roughly 1000 times smaller than that now, i.e., at a redshift $z \equiv \Delta\lambda/\lambda \approx 1000$. At that time the matter is in the form of an ionized plasma at a temperature of 3000 K. Only in the epoch nearest ours does the temperature fall sufficiently so that the recombination of hydrogen occurs. When the plasma is neutral, it becomes transparent to radiation. One cannot "see" the more distant past with the help of electromagnetic waves. Any contention concerning the homogeneity and isotropy of the Universe which follows from the absence of visible fluctuations in the RR temperature on the celestial sphere applies specifically to the epoch of recombination, when $z \approx 1000$. This is the same as $\sim 10^6$ yr from the beginning of the expansion of the Universe (recall that our epoch corresponds to the time 10^{10} years).

We call "observable" that part of the Universe from which electromagnetic waves can reach the observer without being absorbed or scattered en route. In principle, neutrinos, which are very penetrating particles, could enable one to observe an even larger part of the Universe. Particles emitted at the singularity, moving with the speed of light, could reach the observer; the corresponding distance is called the "horizon."

The "observable" radius is large—about 97% of the "horizon" radius. The consequent volume accessible to observations with the help of RR is 90% of the maximum possible volume. Homogeneity on a large scale is thereby established. However, not excluded is the possibility of inhomogeneity in the Universe and of anisotropy of the expansion before the recombination time of $z \approx 1000$, i.e., for times $t < 10^6$ yr from the start of

5. Note too the still unsolved problems of the energy density of gravitational waves of various wavelengths in the Universe and of whether the densities of neutrinos and antineutrinos can be considered equal.

the expansion. What is more, the present-day structure of the Universe, with its separate galaxies and clusters of galaxies, demonstrates the existence of and the requirement for deviations from the ideal picture of homogeneity and isotropy on all scales.

The general picture of an expanding "hot Universe" set forth above is reliably established and is one of the most important achievements of twentieth century science. Now we shall pass to an account of the searches for answers to more subtle questions, such as the problem of the origin of galaxies and the problem of the beginning of the cosmological expansion. Here we confront questions at the forefront of today's research. Not surprisingly, we will speak in a less definite tone and will enumerate various hypotheses.

Producing a thorough analysis of the deviations from the ideal picture of a homogeneous, isotropic Universe is the most important task confronting modern cosmology. In the analysis it is necessary to combine observations and theory. The theory of small perturbations provides a solid foundation for the research. The most general property of a Universe close to the ideal is that one can classify the deviations from the ideal (perturbations) into separate forms ("modes") that evolve independently of one another. Such a theory is of great value in an analysis of the observations.

One can successfully carry out a complete examination of the ideal model by virtue of its utterly exceptional properties which significantly simplify the mathematics. At each instant the ideal model is characterized by several numbers in all (the radius of the Universe, the expansion rate, the density, the temperature, and so forth). The complete evolutionary picture in this approximation is then described by the dependence of these several numbers on time. Consequently, the problem is no more complicated than the problem of the motion of a single particle. This simplification is achieved by limiting the problem to the homogeneous case, or, as is said in mathematics, by considering a "degenerate" case. In distinction, the most general formulation of the problem requires the consideration of functions of at least four variables—time and the three spatial coordinates. In such a formulation the problem is exceptionally complicated. The difficulty is further aggravated by our lack of knowledge of the initial conditions—the conditions at the beginning of the expansion. It is necessary either to investigate the large number of variants of the initial state or, having an (incomplete!) idea of the present-day state, to solve the equations for the behavior in the past. Both approaches encounter great difficulties.

The theory of small perturbations is remarkable in that it combines generality of formulation with mathematical simplicity. On the other hand, by virtue of the small amplitude of the perturbations or, as is said in mathematics, by virtue of the linearity of the problem, an examination of each mode reduces simply to finding several functions of time. In other words, the problem is of the same class of difficulty as the problem of the motion

of a point. This remarkable combination of generality and simplicity explains why a significant part of the book is devoted to a study of the theory of small perturbations of the homogeneous, isotropic Universe.

In the theory of small perturbations, one considers the evolution of separate modes. This is, however, incomplete in the sense that a determination of the absolute value of the amplitude of the various modes is beyond the theory. One must take this quantity from outside the theory—from observations or from other theoretical considerations. The theory tells us, for example, that a density perturbation encompassing a mass greater than $10^{15}\,M_\odot$ grows by a factor of 100 in the time that the scale of the Universe also grows by a factor of 100 (for example, from $t = 10^6$ yr to $t = 10^9$ yr). However, whether or not this kind of perturbation occurs in nature, its amplitude remains unknown in terms of this theory alone. But what do observations tell us? Direct observations of the matter distribution tell us that on scales of the order of 10 Mpc and less, the inhomogeneity is large. This maximum scale is roughly the average distance between neighboring clusters of galaxies. For ρ_0 between 10^{-31} g cm^{-3} and $\rho_c \approx \frac{1}{2} \times 10^{-29}$ g cm^{-3}, the mass in a sphere whose diameter is 10 Mpc lies between $7 \times 10^{11}\,M_\odot$ and $3 \times 10^{13}\,M_\odot$. The density perturbations on this scale today satisfy $(\delta\rho/\rho)_0 \approx 1$. When do such large perturbations form? Apparently in the recent past, at a redshift $z \approx 5$.[6]

Indeed, one can compute, though only very roughly, that the matter density in clusters is of the order of the average density at just that epoch when the clusters-to-be separate from the expanding background and transform into bound configurations. One can clearly imagine that before some definite redshift z_1 (i.e., for $z > z_1$), the matter is expanding, while thereafter (for $z < z_1$), it divides into separate parts—the clusters of galaxies. The distance between these parts continues to increase, but the parts themselves are "preserved." They are maintained in a state of dynamic equilibrium without a change in their average density. Therefore, the present average density of the matter in clusters is roughly characteristic of the average matter density at the epoch of formation. It is this fact which permits one to estimate the formation time.

Using perturbation theory, one can draw from this a conclusion concerning the amplitude of density perturbations and of other quantities in the distant past. For the plasma state in the RD stage, the amplitude $\delta\rho/\rho \approx 10^{-3}$ is obtained for density oscillations. Correspondingly, both the velocity of the separate elements of matter (expressed in units of the velocity of light) with respect to the background of the general expansion and the dimensionless

6. To each value of the redshift z, there corresponds a definite time in the past. This time is at an epoch in the past separated from our epoch by the time interval necessary for the light ray to undergo the corresponding redshift. As an aid to the reader, it is convenient to remember that the quantity $(1 + z)$ shows how many times the scale of the expanding Universe increases from the given epoch to ours.

perturbations of the spacetime metric are of the same order, 10^{-3}. All of this relates to scales characteristic of the present day scale of clusters of galaxies.

Upon the inclusion of theoretical considerations, radio observations of the fluctuations of RR and of its spectrum permit one to estimate the magnitudes of the perturbations as functions of the scale (or mass), i.e., to estimate the spectrum of perturbations. We will discuss this in detail in the appropriate part of the book.

Summarizing, one can say that the picture of the Universe sketched here represents a weakly perturbed (almost homogeneous), expanding Universe with a definite initial (and large) entropy. In support of this picture are the measurements of the spectrum and spatial distribution of the relic radio emission. This is the most important result from the observations and theoretical developments of recent years.

In the picture outlined above, the formation of galaxies and of other structural peculiarities is explained by deviations from homogeneity. Such deviations, though, are sufficiently small that they do not invalidate the desirable features of the homogeneous model. However, such a formulation is not universally accepted. On the one hand, doubts have been raised as to whether this formulation is capable of explaining several observed properties of the Universe. We refer particularly to the observed rotation of galaxies and the presence of galactic-scale magnetic fields. Perhaps one should also add to this list of difficulties associated with the theory of a weakly perturbed homogeneous model the question of the origin of quasars. For this, however, it would be necessary to establish with certainty the nature of quasars—their structure and sources of energy.

Several years ago, the problem of the origin and intensification of magnetic fields on cosmic scales was very troublesome. The properties causing the difficulties are the well-known high electrical conductivity of a plasma and the consequent conservation of magnetic flux. It is the latter effect to which one refers when it is said that magnetic fields are "frozen into" the plasma. Maintenance of a magnetic field requires no external sources of electric current because the decay time of a current in the plasma is on cosmic scales. The problem was that it seemed for a time that this same "frozen in" property would prevent an ordered magnetic field from growing in a plasma originally field free.

In this connection, "magnetic" cosmological models with a primordial field have been examined. Inasmuch as the field picks out a definite direction in space, though, such models are distinguished by primordial anisotropy.

Now, however, it is possible to regard this obstacle to the problem of the origin of galactic magnetic fields as surmounted. Among the three-dimensional motions of the plasma during the formation of rotating galaxies are motions which are capable of generating and intensifying magnetic fields to the values observed today. Hence the hypothesis of a primordial magnetic

field has become unnecessary. The question remains unsettled, however, because a demonstration of the absence of a small initial field (corresponding to a 10^{-10} gauss field after the expansion to the present state) would be very difficult. However, as Landau said, life is short, and one cannot waste it on the study of uninteresting questions that lack great urgency.

The situation regarding rotation is more complicated. The rotation of galaxies is among their most important features, for it is through such rotation, or due to the centrifugal force, that their disklike shape is produced. Moreover, the origin of distinct arms and of spiral structure is a further consequence of the instability of a flat rotating disk. Thus, there is no doubt that rotation exists and plays an important role. A theory ignoring rotation would be quite incomplete and unacceptable.

There is a property of rotational motion analogous to the concept of a frozen-in magnetic field. We refer here to the conservation of angular momentum of an isolated system. If galaxies arise from "potential" or vortex-free (irrotational) perturbations, as is assumed in the theories described above, then it is difficult at first glance to understand the origin of the rotation of galaxies. The simplest explanation of the observed rotation is that elements of the medium present in the RD stage are endowed with primordial rotation. It is in just this way that the hypothesis of a turbulent Universe and of initial photon vortices is formulated. A later development is to use the modern theory of turbulence to obtain more detailed relationships among the angular momentum of a galaxy, the mass of a galaxy, and so forth.

However, one can also explain the origin of the rotation of galaxies within the confines of a vortex-free theory. At the time of the formation of galaxies from nearly homogeneous matter, neighboring protogalaxies interact in the strongest possible way. Thus, the law of angular momentum conservation does not apply to each galaxy individually. It is therefore by no means excluded that rotation can arise spontaneously in the evolution of perturbations which are originally vortex-free.

The heated discussions between the proponents of the "potential" and the proponents of the "vortex" theories relate to two difficulties with the vortex theory. The first is that the theory apparently leads to an excessively early isolation of the galaxies from the expanding matter. This isolation occurs when the average density of the Universe is still rather high—about 10^{-20} g cm^{-3}. This in turn leads to several properties for galaxies that are difficult to reconcile with observations. The second problem is related to the assumption concerning the existence of primordial vortex motions. This assumption requires a change in the character of the early stages of the cosmological expansion itself. Even modest, subsonic vortex motions require a break with the theory of an isotropic expansion and of a weakly perturbed, homogeneous model. According to the revised theory, at least the first 10^5 s of the expansion are not described by the homogeneous, isotropic model.

In addition, by admitting the vortex theory, one must also alter the theory of primordial nucleosynthesis, for this synthesis occurs during the first 100 s of the expansion.

Finally, we cannot fail to mention still another point of view concerning the origin of galaxies. In this theory galaxies form not from the rarefied, expanding gas by means of the gravitational condensation of small inhomogeneities, but by means of explosions of hyperdense bodies. The merit in this point of view is its direct connection with powerful outbursts observed to occur deep in space where galaxies are located. However, the difficulties associated with such a theory are extraordinarily great. Here we mention only that the proponents of this theory must address themselves to the violations of the fundamental laws of physics implied by the existence of pregalactic bodies. For example, the conservation laws for angular momentum and energy appear to be violated in such bodies. To be sure, we do not think that the solution to the problem of galaxy formation lies along a path involving a rejection of the foundations of modern physics.

Earlier we outlined in general terms a picture of the hot, expanding Universe from the first seconds to the time of galaxy formation. We are thus led to a problem of great importance: the question of the very beginning of the cosmological expansion. From what does the expansion begin? How does the Universe expand at the very beginning? What happens before the start of the observable expansion?

Consider the earliest stages of the expansion, the stages before the synthesis of the chemical elements. This means times less than 1 s from the beginning of the expansion. Perhaps the behavior of these stages actually differs sharply from that of the homogeneous and isotropic model and is characterized, for example, by a markedly anisotropic expansion. Such anisotropic expansion may well be that described by the anisotropic models that have been intensively studied in the last few years. Quite recently, in fact, the question was discussed of whether there is actually a singularity—of whether there is actually infinite density at the start of the expansion. Is it possible that a singularity is characteristic only of the ideal homogeneous, isotropic model, and that for a general, asymmetric, nondegenerate solution, contraction of the matter in some preceding epoch changes into expansion without the occurrence of a singularity? What we know now is that there is strict proof of the existence of a singularity in the past even if the behavior of the early stages differs sharply from the presumed homogeneous, isotropic expansion. We will discuss these problems in detail in the appropriate places in the book. Here we merely want to emphasize one thing and shall do so by posing a question. Suppose we examine (1) the various types of cosmological models laying some claim to a description of the Universe close to the start of the expansion and (2) the specific theories of various types of perturbations possibly existing at the very early stages. This examination would lead us to a very fundamental question: Given the data of all possible

present-day observations and using the laws of physics, can one unambiguously establish the past history of the Universe, viz., the behavior of the very early stages of expansion, of the singularity itself, and of the epoch "before the expansion" (in the case that the last phrase has any meaning)?

To approach this question, think of thermodynamics. It is easy to imagine that under certain conditions a wide class of initial states quickly leads to the same final state, which in turn serves as an initial state for further evolution. The actual initial state is thus *forgotten.* In this spirit, we perceive a new direction in cosmology—theoretical research of the unobserved early stages. It therefore becomes a question of finding that cosmological model which can be obtained from a wide class of initial early states.

It turns out, for example, that a whole group of anisotropic initial states leads to isotropic expansion. Is this the explanation of why the Universe expands isotropically? Perhaps this means that, of a large number of *in principle possible initial Universes* (a blasphemous phrase!), many become identical to the Universe observed and studied by us. But there is only one Universe, and therefore statistical arguments do not seem applicable.

Along with the question of why the Universe expands homogeneously and isotropically, several other questions also confront cosmology. Why is the entropy of the Universe large? Why is the Universe hot at the start of the expansion? Finally, why are the perturbations of such magnitude that they lead to galaxy formation in our epoch and not much earlier or much later? It is possible that a new theoretical formulation of these questions will arise from a consideration of quantum cosmology.

In conclusion we emphasize the basis—the physical laws upon which theoretical cosmology is constructed. It would be tempting to proclaim that new laws are necessary to describe phenomena on the scale of the Universe—similar to what occurred when quantum laws were developed to describe the microcosm. However, such an analogy is superficial and oversimplified.

The microcosm called attention to the insufficiency of classical theory 50 years before the creation of quantum theory. In cosmology, however, we have not yet encountered any insoluble contradictions between theory and experiment or any internal logical difficulty in the theory. The theory of the expanding Universe was created before the actual discovery of galaxy recession. Similarly, the theory of the hot Universe was created long before the discovery of the RR. These factors support the correctness of our theories and the applicability of them to scales characteristic of the Universe. It is necessary only to emphasize that the question is not whether new phenomena should be excluded. Indeed, extraordinary and seemingly new phenomena arise in cosmology as a result of the application of existing theory.

GTR, for example, leads to the new situation of relativistic collapse. New phenomena also arise from combining the quantum theory of matter with the theory of gravity. Theories of particle creation have repeatedly been

put forward. In such theories, the particle creation occurs throughout space and time—in violation of the laboratory laws of physics. One can have respect for this direction of research as poetic foresight, but not as scientific prediction. This is because such theories of continuous matter creation are refuted by observations. However, there is a period in the evolution—close to the singularity—when particles are created in the then very strong gravitational field. In this case, though, it turns out that intense particle creation proceeds only when the expansion is anisotropic—a result which is by no means obvious.

Finally, new phenomena may arise in this same period when account is taken of the quantization of the metric. This is a possible theoretical development applicable to conditions definitely unknown to modern physics.

Thus, the fundamental propositions of modern physics (based on experiments and self-consistency) plus the application of them to the new conditions at the singularity itself plus the observational data obtained with the help of both terrestrial instruments as well as experiments in space form the foundation upon which modern cosmology is, or must be, based.

In conclusion, we want to make a historical remark. It is well known that A. Friedmann created the theory of the evolving Universe in 1922–1924. The famous discussion with Einstein on the pages of the *Zeitschrift für Physik* called attention to Friedmann's work. We therefore cannot agree that Lemaitre (1927) "independently" established the laws of the expanding Universe.

After Hubble's discovery (1929) of the linear redshift-distance relation, the mathematical solutions attained the status of established theory. In reviewing the situation in 1931, Einstein wrote: "Friedmann was the first to follow this way . . ." (Einstein 1966; the reference is to the "Collection of Scientific Papers of A. Einstein" published by the Nauka publishers).

I THE HOMOGENEOUS, ISOTROPIC UNIVERSE

ITS EXPANSION AND GEOMETRICAL STRUCTURE

1 LOCAL PROPERTIES OF THE HOMOGENEOUS, ISOTROPIC COSMOLOGICAL MODEL

1.1 The Local Law for the Velocity Distribution

The homogeneous, isotropic cosmological model represents an idealized picture of the Universe in which at any given instant of time: (*a*) at different spatial points all observable quantities are the same, and (*b*) at any spatial point all directions are equivalent. The first property is homogeneity and the second isotropy.

We shall be studying the homogeneous and isotropic cosmological model. We will show that if the distribution and motion of matter is homogeneous and isotropic at any moment, then this property is preserved for the entire evolution of the system. It is then sufficient to follow a single volume element of matter; the fate of all other volume elements will be exactly the same.

For an analysis of the evolution of such a model, it is convenient that the Newtonian theory of gravity is applicable. Why is this so? The model under consideration is a particular case of the spherically symmetric solution, in which any point can be chosen to be the origin. Now it is well known that in the classical, Newtonian theory of gravity, a spherically symmetric distribution of matter does not give rise to a gravitational field inside a spherical cavity (of radius R) concentric to the matter distribution. In fact, this is also valid in GTR if the Universe is homogeneous and expands isotropically. (Both contentions are examined in detail in Vol. 1.) Thus, consider a sufficiently small sphere of radius R. The gravitational field due to its own mass will be weak, and the velocities of relative motion inside the sphere will be small. Hence, one can use Newtonian theory to describe the evolution. Under these circumstances, which describe our Universe quite well, the use of GTR allows one to prove that GTR itself is unnecessary for the solution of local problems in cosmology.

We will use the Newtonian description (McCrea and Milne 1934; Milne 1934, 1935; Zel'dovich 1963; Callan, Dicke, and Peebles 1965) to derive formulae describing the evolution of the model. Of course, when we pass from the study of local properties to that of global properties, it will be necessary to return to GTR. Thus, in a small region, we shall apply Newton's theory. Of course, the same formulae we obtain could also be obtained directly from the equations of GTR; we will do this in the usual way in

§ 2.1. However, the meaning of the formulae would not be so evident if we were to use only the latter method.

The motion of matter will be examined in a coordinate system chosen so that the matter at the origin of the coordinates is at rest. In this system, the matter located some distance from the origin moves. Let the velocity be positive, i.e., be directed away from the origin of coordinates, and let it be proportional to the distance from the origin. In vector form, such a law for the velocity distribution is written

$$v = Hr \tag{1.1.1}$$

with $H > 0$. H is called the Hubble parameter and is independent of the magnitude and direction of the vector r. However, H does depend on time, though this dependence will not be examined in detail until § 1.2. It is evident that this velocity distribution is isotropic: to an observer located at the origin of coordinates, no direction has special significance. But at an arbitrary point A, described by the radius vector r_A, the matter moves with velocity v_A, so it might appear that the isotropy is destroyed. To show that this is not the case, pass to the system of coordinates whose origin is at the point A and moves with the velocity v_A. In other words, translate the origin of coordinates and make a Galilean transformation to the moving system. We will then denote by a prime quantities measured in the new system. It is obvious that $r' = r - r_A$, so

$$v' = v - v_A = Hr - Hr_A = H(r - r_A) = Hr' . \tag{1.1.2}$$

In the new system, consequently, the same law for the velocity distribution holds: v' depends on r' in the same way that v depends on r; the isotropy is thus maintained.

Remarkably, the velocity distribution (1.1.1) is precisely the one which does not select out any particular point. Only this type of velocity distribution is isotropic and homogeneous. An observer moving along with the matter at any point sees all neighboring particles receding.

The law of expansion (1.1.1) leads to the following picture of the local expansion. The distance between any pair of material points A and B changes with time according to the law $dr_{AB}/dt = Hr_{AB}$, from which it follows that

$$r_{AB}(t) = r_{AB}(t_0) \exp \int_{t_0}^{t} H(t)dt . \tag{1.1.3}$$

Now consider the law for the evolution of density. We will take a sphere containing a fixed mass M and denote its radius by $R = R(t)$. The matter density is $\rho = M/[(4\pi/3)R^3]$, from which we find

$$\frac{d\rho}{dt} = \frac{-3M}{(4\pi/3)R^4} \frac{dR}{dt} .$$

In this formula we substitute $dR/dt = v = HR$ and obtain

$$\frac{d\rho}{dt} = -3\rho H . \tag{1.1.4}$$

One can also obtain this equation more formally from the equation of continuity

$$\frac{\partial \rho}{\partial t} = -\operatorname{div}(\rho v) .$$

Assume that ρ does depend on time but does not depend on the spatial coordinates, and substitute for v according to formula (1.1.1):

$$\operatorname{div}(v) = \operatorname{div}(Hr) = 3H .$$

From this it again follows that

$$\frac{\partial \rho}{\partial t} = -3H\rho . \tag{1.1.5}$$

Moreover, $\partial\rho/\partial t$ does not depend on the spatial coordinates. Consequently, if at some moment ρ does not depend on the spatial coordinates, then by the law of expansion (1.1.1), it is also independent of the spatial coordinates at all succeeding times; therefore, $d\rho/dt = \partial\rho/\partial t$. Thus, homogeneity, specified at an initial time, is a preserved property.

The same result also follows for the properties of the velocity distribution; i.e., they are preserved as time evolves. In the rest system of the center of a sphere, $v = Hr$. Let us find the law which governs the change of velocity. The acceleration at a given point r inside the sphere depends on the gravitational force caused by that mass located within the interior sphere, i.e., the sphere with radius $r = |r|$. We shall call this mass $M(r)$, where $M = (4\pi/3)\rho r^3$. The acceleration g quantitatively is $GM/r^2 = (4\pi/3)G\rho r$ and is directed to the center of the sphere, so that, in vector form,

$$g = -\frac{4\pi}{3} G\rho r .$$

Consequently, $v(t + \Delta t) = v(t) + \Delta t g$. Inasmuch as $v(t) \sim r$ and $g \sim r$, the velocity is proportional to r at any preceding or subsequent time. Hence $v = Hr$ always holds, although the quantity H does depend on time.

In conclusion, we emphasize that the above remarks remain a valid description of matter even when pressure is present. Indeed, by virtue of the homogeneity, there is no gradient of the pressure or any other force connected with the pressure that influences the motion. It is necessary to say, though, that there exists a relativistic effect whereby pressure creates an additional gravitational field. This effect is substantial only in the case when the equation of state of the matter is relativistic and when the pressure P is of the order of the energy density: $P \approx \epsilon$. We will discuss this effect further in § 1.5. In the following sections, § 1.2–4, we shall continue to

assume that the pressure is much less than the energy density: $P \ll \epsilon$. Indeed, for the present epoch this condition is very well satisfied.

1.2 The Law of Evolution and the Critical Density

Let us again examine a sphere which contains a mass M and find the acceleration of a particle located at the surface of the sphere. A force of attraction acts on the particle, producing an acceleration

$$\frac{d^2R}{dt^2} = -\frac{GM}{R^2} . \tag{1.2.1}$$

We substitute into equation (1.2.1)

$$\frac{dR}{dt} = HR , \qquad M = \frac{4\pi}{3}\rho R^3$$

and obtain

$$\frac{d(HR)}{dt} = R\frac{dH}{dt} + H\frac{dR}{dt} = R\frac{dH}{dt} + RH^2 = -\frac{4\pi}{3}G\rho R ,$$

or

$$\frac{dH}{dt} = -H^2 - \frac{4\pi}{3}G\rho . \tag{1.2.2}$$

Equations (1.1.4) and (1.2.2) form a system completely determining the changes with time (both in the past and in the future) of all local properties of the Universe. Just as expected, the radius R of the arbitrarily chosen mass M does not appear in these equations. However, in order to demonstrate the manifestly general character of the solution, it is convenient to make use of both quantities M and R in the subsequent development.

After multiplying by dR/dt, the equation of motion (1.2.1) can be integrated to give

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 - \frac{GM}{R} = \mathsf{A} . \tag{1.2.3}$$

This equation has the form of an energy-conservation law. The first term on the left is the kinetic energy of a unit mass; the second term on the left (negative) is the potential energy of that mass; the constant A is the total energy of a unit mass at the surface of the sphere.

With t_0 the present time, consider known the values $H_0 \equiv H(t_0)$ and $\rho_0 \equiv \rho(t_0)$ and select a value R_0 for the radius of the sphere; then, $M = (4\pi/3)R_0^3\rho_0$. We know that $(dR/dt)_{t=t_0} = H_0R_0$, and thus the value of the constant on the right side of equation (1.2.3) can be determined. We obtain

$$\mathsf{A} = \frac{1}{2}\left(\frac{dR}{dt}\right)^2_{t=t_0} - G\frac{4\pi}{3}\frac{\rho_0 R_0^3}{R_0} = \tfrac{1}{2}(H_0R_0)^2 - G\frac{4\pi}{3}\rho_0 R_0^2 .$$

It follows from equation (1.2.3) that

$$\left(\frac{dR}{dt}\right)^2 = \frac{8\pi}{3}\frac{G\rho_0 R_0^3}{R} - \frac{8\pi}{3}GR_0^2\left[\rho_0 - \frac{3H_0^2}{8\pi G}\right]. \tag{1.2.4}$$

This equation can be integrated, and we will present the solution in § 2.1. But it is highly instructive to give a qualitative analysis of the solution now.

The general character of the solution to equation (1.2.4) is easy to perceive. At the present time dR/dt is positive. Consequently, compared to their present values, R is smaller in the past, while $(8\pi/3)G\rho_0 R_0^3/R$ is larger. Thus, dR/dt is necessarily larger in the past than now; hence there is a time when

$$R = 0\,, \qquad \frac{dR}{dt} = +\infty\,.$$

A prediction of the future depends on the sign of the bracketed quantity $[\rho_0 - 3H_0^2/8\pi G]$ in equation (1.2.4). We define the quantity ρ_c by

$$\rho_c \equiv \frac{3}{8\pi}\frac{H_0^2}{G} \tag{1.2.5}$$

and denote by Ω_0 the ratio ρ_0/ρ_c:

$$\Omega_0 \equiv \rho_0/\rho_c = \frac{8\pi}{3}\frac{G\rho_0}{H_0^2}\,.$$

We note here that a closely related quantity, q_0, satisfying $q_0 = \frac{1}{2}\Omega_0$, is also widely used in cosmology and is called the deceleration parameter. This name arises from the defining expression

$$q_0 \equiv \left[\frac{1}{H^2}\left(-\frac{1}{R}\frac{d^2R}{dt^2}\right)\right]_{t=t_0}. \tag{1.2.6}$$

The relation $q_0 = \Omega_0/2$ can then be easily verified by using equations (1.2.1), (1.2.4), and (1.2.5).

If $\rho_0 > \rho_c$, the expression in brackets in equation (1.2.4) is positive. Then, as R increases, a value will be achieved which makes the entire right side of that equation vanish. At this time, the expansion will cease and contraction will ensue (Fig. 1).

If $\rho_0 < \rho_c$, the expansion will continue without bound. In the limit when $t \to \infty$, $R \to \infty$, we then have

$$\frac{dR}{dt} = \left[\frac{8\pi}{3}GR_0^2(\rho_c - \rho_0)\right]^{1/2} = \text{const}\,.$$

The corresponding curve is shown in Figure 2.

As was already noted, the distance between any pair of objects changes in proportion to R. The distance passes through a maximum at a time t_m in the case of Figure 1 and increases without bound in the case of Figure 2. In the first case ($\rho_0 > \rho_c$) Hubble's parameter decreases from H_0 at $t = t_0$

to $H = 0$ at $t = t_m$ and then to $H \to -\infty$ at $t \to t_{\infty}$. In the second case ($\rho_0 < \rho_c$), H decreases, but always remains positive; in the limit as $t \to \infty$, $dR/dt = V_{\infty}$, $R = V_{\infty}t$, and $H = t^{-1}$. These results are asymptotically exact.

Which case actually occurs? Is $\rho_0 > \rho_c$ or is $\rho_0 < \rho_c$? For an answer it is evidently necessary to know the present-day values of ρ_0 and H_0.

We will discuss in detail the status of the question of determining ρ_0 and H_0 from observations in § 3.9. In what follows we shall write $H_0 = 100h_0$ km s^{-1} Mpc^{-1}. The most probable value for H_0 at present (Sandage and

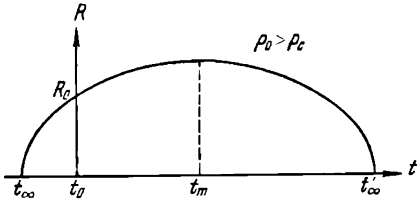


FIG. 1.—Time evolution of the distance R between two points when the density ρ_0 is greater than the critical density ρ_c. The time t_0 represents the present, t_{∞} and t'_{∞} are the times when the density is infinite, and t_m is the time of maximum expansion.

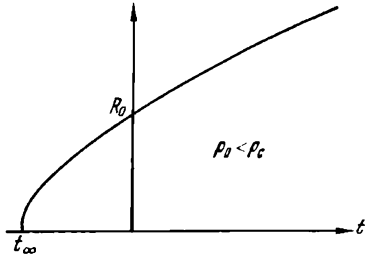


FIG. 2.—Time evolution of the distance between two points when the density ρ_0 is less than the critical density ρ_c.

Tammann 1974*a*, *b*, *c*, *d*, 1975*a*, *b*) is $H_0 \approx 55$ km s^{-1} Mpc^{-1}, $h_0 \approx \frac{1}{2}$. We then have $H_0^{-1} \approx 3 \times 10^{17}\, h_0^{-1}$ sec $\approx 10^{10}\, h_0^{-1}$ yr. The corresponding value of the critical density is $\rho_c \approx 2 \times 10^{-29}\, h_0^2$ g cm^{-3}. Meanwhile, it has been established sufficiently reliably that the average density of material in the Universe, ρ_0, is not less than $\rho_0 = 3 \times 10^{-31} h_0^2$ g cm^{-3}. This limit to ρ_0 is determined by the mass of visible matter in galaxies but does not take account of the mass of possible intergalactic material. If there is not a significant density of material between galaxies, then $\rho_0 < \rho_c$ and the case shown in Figure 2 is realized.

However, it is not excluded that in fact the density of matter is greater—in particular due to intergalactic ionized hydrogen or other difficult-to-observe forms of material (see the chapters in Part III for more detail).

As a conclusion to this section, we emphasize the following important observation. We have considered the evolution of a volume element of matter in the Universe under the influence of the force of gravity. At the same time it was assumed that the cosmological constant Λ in the gravitational equation is equal to zero. A discussion of the meaning and importance of the cosmological constant, and of its possible nonzero value, appears in § 1.1.2. Note that a nonzero Λ can affect the late stages of expansion. However, as we already emphasized in the Introduction, at the present time there are neither theoretical reasons nor convincing observational data in favor of a nonzero Λ. Therefore, we will consider $\Lambda = 0$ for the duration of this book, with the exclusion of § 4.1.2, wherein are examined in detail the consequences of a nonzero Λ.

1.3 The Duration of the Expansion

Let us find the moment in the past, defined to be $t = 0$, when $R = 0$, i.e., when the matter density is infinite. If the velocity of expansion were constant and equal to the present-day value, then one would be able to write

$$R_0 = \left(\frac{dR}{dt}\right)_{t=t_0} t_0 = H_0 R_0 t_0 . \tag{1.3.1}$$

For the "age" of the homogeneous model of the Universe, i.e., for the elapsed time from $t = 0$ to the present, $t = t_0$, we would obtain

$$t_0 = H_0^{-1} = \frac{3 \times 10^{17}}{h_0} \text{ s} \approx \frac{10^{10}}{h_0} \text{ yr} .$$

As is apparent from equation (1.2.3) and Figures 1 and 2, dR/dt is greater in the past than it is today. Indeed, it is not difficult to show, using the formulae of the previous section (see eq. [1.2.4]), that

$$t_0 = \frac{1}{H_0} f(\rho_0/\rho_c) = \frac{1}{H_0} f(\Omega_0) . \tag{1.3.2}$$

The analytic form of the function $f(\Omega_0)$ is presented in the Appendix to this section. A plot of $f(\Omega_0)$ is given in Figure 3. The dashed curve describes the case when photons and neutrinos dominate the density (see § 1.7).

1.4 Two Particular Solutions; the Early Stage

The solution for the case $\rho_0 = \rho_c$, $\Omega_0 = 1$, has an especially simple form. Equation (1.2.4) then reduces to

$$\left(\frac{dR}{dt}\right)^2 = \frac{8\pi}{3} \frac{G\rho_0 R_0^3}{R} . \tag{1.4.1}$$

The solution of this equation gives

$$R = R_0 \left(\frac{t}{t_0}\right)^{2/3} ; \tag{1.4.2}$$

$$t_0 = \frac{2}{3}\frac{1}{H_0} ; \tag{1.4.3}$$

$$\rho = (6\pi G t^2)^{-1} . \tag{1.4.4}$$

In equation (1.2.4), the second term on the right disappears for $\rho_0 = \rho_c$. But for any present-day $\rho_0 \neq \rho_c$, there is a period in the past when R is so small that one can neglect the constant second term on the right side of equation (1.2.4) in comparison with the first term, which is proportional to

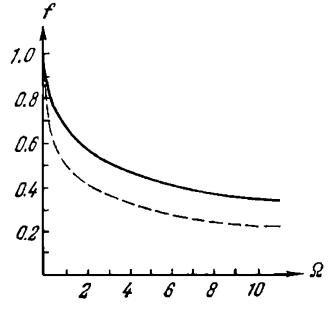


FIG. 3.—The dimensionless age of the Universe (the age in units of the inverse of the Hubble parameter, H_0^{-1}) as a function of the dimensionless density (the density in units of the critical density). The solid line is for a universe filled with zero-pressure material, $P = 0$; the dashed line is for a universe filled with material for which the equation of state is $P = \epsilon/3$.

R^{-1}. Therefore the expression (1.4.4) for the density is universal for the early stages, independent of the value of ρ_0/ρ_c today.

The second special solution to equation (1.2.4) is the one with vanishingly small density, $\rho_0 \ll \rho_c$. In this case, neglecting the term with ρ_0, we obtain

$$\frac{dR}{dt} = H_0 R_0 = \text{const.} , \tag{1.4.5}$$

so that

$$R = R_0 + H_0 R_0 (t - t_0) . \tag{1.4.6}$$

Once again recall that we are neglecting the pressure of matter. Near the singularity, $\rho = \infty$, one may not do this. For the effect of pressure on the law of expansion, see § 1.5.

In this low density approximation

$$t_0 = \frac{1}{H_0} ; \quad \frac{R}{R_0} = H_0 t ; \tag{1.4.7a}$$

$$H = H(t) = \frac{1}{t} = \frac{H_0}{1 + H_0(t - t_0)} .$$

We find for the density

$$\rho(t) = \rho_0 \frac{R_0^3}{R^3} = \rho_0 \left(\frac{t_0}{t}\right)^3 . \qquad (1.4.7b)$$

Let us compare the density in the past, $\rho(t)$, for $t < t_0$, with the critical density in the past $\rho_c(t)$, i.e., with the quantity

$$\rho_c(t) = \frac{3H^2(t)}{8\pi G} .$$

For small $t \ll t_0$, the density $\rho(t)$ grows faster than $\rho_c(t)$. Therefore, for any small but nonzero present-day ρ_0 there is, in the past, a period when the density is close to the critical density $\rho_c(t)$ [calculated using the appropriate

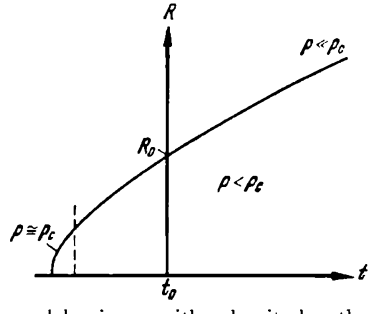


Fig. 4. Expansion of a model universe with a density less than the critical density. To the left of the dashed line is the epoch when the density is close to critical. The density becomes increasingly smaller than the critical density over the course of the expansion, so that $\rho \ll \rho_c$ holds for large values of time t.

instantaneous value $H = H(t)$]. Note that, previously, we meant by ρ_c the present-day value calculated using H_0; this meaning will apply in the rest of the text. The solution of equation (1.2.4) for $\rho_0 < \rho_c$; $\Omega_0 < 1$ consists of two parts (see Fig. 4):

$$\rho = \frac{1}{6\pi G t^2} \quad \text{for} \quad \rho > \rho_c^3/\rho_0^2 , \quad t < 2\Omega_0/3H_0 ; \qquad (1.4.8a)$$

$$\rho = \rho_0(t_0/t)^3 \quad \text{for} \quad \rho < \rho_c^3/\rho_0^2, \quad t > 2\Omega_0/3H_0 ; \qquad (1.4.8b)$$

the time t_0 in this case is given by the expression

$$t_0 = \frac{H_0^{-1}}{(1-\Omega_0)}\left[1 - \frac{\Omega_0/2}{(1-\Omega_0)^{1/2}} \cosh^{-1}\left(\frac{2}{\Omega_0} - 1\right)\right] . \qquad (1.4.9)$$

Remember that, in solution (1.4.8), the symbol ρ_c means the quantity calculated according to formula (1.2.5) with the present-day value of Hubble's parameter, $H_0 = H(t_0)$. Correspondingly, $\Omega_0 = \rho_0/\rho_c = 8\pi G\rho_0/3H_0^2$ is calculated using the present-day values H_0 and ρ_0.

1.5 The Effect of Pressure on the Expansion Law: Qualitative Considerations

In the preceding sections, we examined the motion of matter whose pressure can be neglected. Examples of such matter are gas at a moderate temperature or dust that is uniformly distributed in space. The totality of galaxies may be considered as "dust." However, as we shall see later, those particles composing the principal part of the mass density at the earliest stage of expansion of the Universe are photons and neutrinos (and perhaps gravitons), i.e., particles moving at the speed of light.

For any kind of matter, the mass density ρ (in units of g cm^{-3}) can be connected with an energy density ϵ (in units of ergs cm^{-3}) by the relativistic relation

$$\rho = \epsilon/c^2 . \tag{1.5.1}$$

Now for "ordinary matter" one can neglect the pressure in comparison to the energy density ϵ. But for neutrinos and photons, the pressure is comparable to ϵ:

$$P = \tfrac{1}{3}\epsilon = \tfrac{1}{3}\rho c^2 . \tag{1.5.2}$$

The effect of pressure on the expansion law is most important at the beginning of the evolution. At the present time there are neutrinos and photons which have existed since the beginning of the expansion (for more detail, see chapters 5–8). From this it follows (as we shall show) that in the past, during the early stages, the energy density in neutrinos and photons is significantly higher than the energy density in ordinary matter. However, even if the density of neutrinos and photons is negligibly small and the temperature equal to zero, consideration of pressure would still be necessary at the beginning, when the density of baryons is very high.

In the epoch characterized by densities in excess of nuclear densities, of order $\rho \approx 10^{14}$ g cm^{-3}, even ordinary matter consisting of neutrons, protons, and electrons has a pressure of order $\epsilon = \rho c^2$. The existence of high pressures when the density is high is a result of two effects. First, neutrons, protons, and electrons obey Fermi statistics; at high density, therefore, their average velocity approaches the velocity of light even when the temperature is zero. Consequently, if interactions are ignored, the pressure asymptotically approaches, but remains less than, $P = \epsilon/3$. In addition, the mutual repulsion of nucleons at small separation can, in principle, lead to pressures $P = \epsilon$. For a discussion of possible deviations, see § 6.4.

Let us examine a volume element containing a definite amount of ordinary matter and a definite energy density of neutrinos and photons. In the course of the expansion, the energy density of the ordinary matter decreases inversely with the increasing volume, i.e., as V^{-1} or R^{-3}. Meanwhile, the energy density of neutrinos and photons decreases more rapidly, as $V^{-4/3}$ or R^{-4}. This more rapid decrease is a consequence of the pressure of the neutrinos and photons. Exerting pressure, they perform work on neighboring

volume elements during the expansion: $d(\epsilon V) = -PdV$. Indeed, this is just the law describing the adiabatic expansion of a gas. At the same time, the temperature and energy of the neutrinos and photons vary as $V^{-1/3} = R^{-1}$. That is, photons and neutrinos travel along straight lines and, with respect to the receding matter, they lose energy, or undergo a redshift ("cool off"). For detailed calculations of this effect, based on a study of the equation of motion for freely propagating photons, see § 3.11. If, in a relativistic gas, collisions are neglected, then these conclusions are still valid. Thus $\rho \sim \epsilon \sim P \sim V^{-4/3}$ for a relativistic gas, while $\rho \sim \epsilon \sim V^{-1}$ for ordinary matter. Consequently, near the beginning of the expansion, the density depends primarily on the properties of a relativistic gas and so the pressure is then of the same order as the energy density.

We will now return to the law of expansion for a homogeneous model of the Universe filled with relativistic gas. Can one say that the high pressure is the cause of the expansion of the Universe—that the highly compressed matter expands for the same reason that the molecules of a high-pressure gas, formed by the detonation of an explosive charge, fly apart? No, such a point of view is completely incorrect! The qualitative difference is that the exploding matter is surrounded by air at atmospheric pressure. The expansion results from the difference between the extraordinarily high pressure of the gas (the product of the explosion) and the comparatively weak pressure of the surrounding air. But for a homogeneous Universe, the pressure distribution is strictly homogeneous! Consequently, there are no pressure differences between neighboring regions at a given instant; hence there is no force which could affect the expansion, let alone be the cause of the expansion. In existing theory, the expansion itself is a result of an initial velocity distribution. The relativistic effect of pressure on the expansion law is subtle. Indeed, as will be shown below, it is an effect of the opposite sign—it slows, rather than accelerates, the expansion.

This relativistic effect is connected with two factors. The first of them, described above, is the change in energy which accompanies a change in the volume occupied by a definite number of particles in the presence of pressure. As the expansion proceeds, the matter performs work and its energy decreases. The energy density of a relativistic gas is proportional to R^{-4}, not to R^{-3} as is the case for dust. Therefore, the mass-energy of an expanding sphere, such as the one we examined, is not constant for the case of a relativistic gas, but is greater during the earlier stages of expansion.

The second factor is the creation (according to GTR) of a gravitational field not only by mass density, but also by all other components of the stress-energy tensor T_{ik}. For the case of a relativistic gas, the components of T_{ik} describing the pressure satisfy $T_1^1 = T_2^2 = T_3^3 = -P$ and are of the same order as that describing the energy density $T_0^0 = \epsilon$ since $\epsilon = 3P$. In this case both the pressure and the density are important in creating the gravitational field. The gravitational field caused by a mass element of a

relativistic gas is twice as strong as that caused by a mass element with the same energy density but without pressure.

The formulae describing this effect, naturally, can be consistently deduced only by using GTR, since the effect is relativistic. In the following section, however, before deducing the formulae, we will make use of simple relations which follow from GTR (and which will be determined later; see § 2.1), in order to examine qualitatively the effect of pressure on the expansion law.

1.6 The Equation of Motion

With allowance for pressure the attractive acceleration (deceleration) (calculated from GTR by Tolman 1930; Whittaker 1955; Pauli 1958; see also eq. [3.3.5] of Vol. 1) is

$$g = -\frac{4\pi GR^3}{3R^2}\left(\rho + \frac{3P}{c^2}\right) = -\frac{4\pi G}{3c^2}R(\epsilon + 3P) = \frac{d^2R}{dt^2}. \quad (1.6.1)$$

We are considering a sphere with a small radius and with an expansion velocity that is small compared to the velocity of light. Under these conditions, expression (1.6.1) remains valid.

McCrea (1951) used the above expression to show that it is simple to obtain the law governing changes in the density and the distance if one uses Newtonian mechanics and the Newtonian theory of gravity and replaces ρ by $\rho + 3P/c^2$. His results are reproduced below.

Let us again consider a sphere of radius R. The equation of motion has the form (1.6.1). This equation is given in the book of Pauli (1958), and we will derive it using GTR in § 2.1. The total number of conserved particles (nucleons) inside the given sphere is constant. Therefore the density of conserved particles,

$$n = \frac{N}{(4\pi/3)R^3}, \quad (1.6.2)$$

depends on the radius, as it did formerly. However, the dependence of the energy density on the radius is more complicated. It satisfies the equation

$$dE = d\left(\frac{4\pi}{3}R^3\epsilon\right) = -PdV = -P\cdot 4\pi R^2 dR. \quad (1.6.3)$$

Remarkably, the set of equations (1.6.1) and (1.6.3) leads to an integral of the motion which takes a very simple and obvious form similar to that of equation (1.2.3). We multiply equation (1.6.1) by dR/dt:

$$\frac{d^2R}{dt^2}\left(\frac{dR}{dt}\right) = \frac{1}{2}\frac{d}{dt}\left(\frac{dR}{dt}\right)^2 = -\frac{G}{c^2}\left[\frac{4\pi}{3}\epsilon R\frac{dR}{dt} + 4\pi PR\frac{dR}{dt}\right].$$

Then, using equation (1.6.3), we transform the expression in brackets to

$$\frac{4\pi}{3}\epsilon R\frac{dR}{dt} + 4\pi PR\frac{dR}{dt} = -\frac{4\pi}{3}\frac{d}{dt}(\epsilon R^2).$$

Finally we obtain

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 - G\frac{1}{R}\left(\frac{4\pi}{3}R^3\frac{\epsilon}{c^2}\right) = \text{const}. \qquad (1.6.4)$$

Thus, taking account of the connection between pressure and energy density, we find that $\rho = \epsilon/c^2$ enters the term in expression (1.6.4) which describes gravity (the term proportional to G) even though $\rho + 3P/c^2 = (\epsilon + 3P)/c^2$ enters the expression for the acceleration. For a determination of the constant on the right side of equation (1.6.4), it is necessary to substitute the present value of the various quantities and to use $dR/dt = HR$. The result obtained is exactly like that obtained earlier: the sign of the constant depends on the relationship between the actual present density, $\rho_0 = \epsilon_0/c^2$, and the critical density ρ_c. Notice, however, that if a large part of the matter consists of particles moving with the velocity of light, so $P = \epsilon/3$, then

$$\frac{\epsilon + 3P}{c^2} = \frac{2\epsilon}{c^2} = 2\rho .$$

For the same density, therefore, the deceleration is twice what it would be in the absence of pressure.

For reference, we note that the critical density $\rho_c \approx 2 \times 10^{-29} h_0^2$ g cm^{-3} corresponds to radiation with a blackbody temperature of $40\, h_0^{1/2}$ K or to a Fermi distribution of neutrinos (Fermi gas at zero temperature) with a Fermi energy $7.5 \times 10^{-3} h_0^{1/2}$ eV and a number density of $2 \times 10^7 h_0^{3/2}$ cm^{-3}. Data relating to the actual density of electromagnetic radiation and of neutrinos in the Universe will be examined later.

1.7 The Expansion Time

Taking the density of ordinary matter to be small, let us assume that the density of neutrinos and photons is significantly larger. Then it is easy to integrate equations (1.6.3) and (1.6.4) if one sets $P = \epsilon/3$. The result of the integration is given at the end of § 2.1. Here we will present the results for the age of the Universe:

$$t_0 = \frac{1}{H_0}\left(\frac{1}{1 + \Omega_0^{1/2}}\right), \qquad (1.7.1)$$

where $\Omega_0 = \rho_0/\rho_c$. The expression is valid for $\Omega_0 > 1$ as well as for $\Omega_0 < 1$. The function $f = (1 + \Omega_0^{1/2})^{-1}$ relevant for the case $P = \epsilon/3$ is plotted as the dashed line in Figure 3. This result allows one to obtain rough but reliable estimates of the upper limit on the possible neutrino density (Peres 1960; Smorodinsky and Pontecorvo 1961; Weinberg 1962; Zel'dovich and Smorodinsky 1961).

Indeed, since the age of the Earth, that of the solar system, and that of the chemical elements is of order $(4\text{–}5) \times 10^9$ yr, then one must find that

t_0 is greater than this quantity. Substituting $1/H_0 \approx 10^{10}/h_0$ yr, we obtain the limit $\Omega_0 h_0^2 \lesssim 4$. This condition applies without distinction to nonrelativistic particles with nonzero rest mass $m \neq 0$; to invisible galaxies or collapsed stars; and to gravitational waves, neutrinos, or other weakly interacting particles with zero rest mass.

For neutrinos in thermal equilibrium, we find $T_\nu < 50$ K. If one assumes that the neutrino distribution obeys the Fermi law, then the Fermi energy is less than 0.01 eV. Such neutrinos are impossible to observe using the methods of contemporary nuclear physics, but their gravitational action on a cosmic scale provides a much more sensitive means of detection. We may indicate a still more sensitive means of detection. In a later chapter it will be shown that the gravitational effect of particles with zero rest mass on

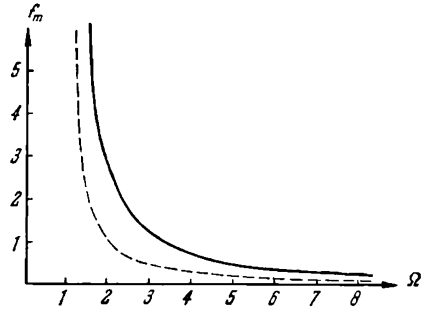


FIG. 5.—The dimensionless time required for the Universe to reach its maximum radius, as a function of its density. The solid line is for $P = 0$, while the dashed line is for $P = \epsilon/3$.

the expansion of the Universe markedly changes the reactions of chemical-element synthesis. Under several natural assumptions (although not as unquestionable as the arguments in the text), this effect leads, upon comparison with observations, to a still more stringent present-day limit: $\rho_0(m = 0) < 10^{-32}$ g cm^{-3}. (For more detail, see chap. 7.)

In essence, the above methods were suggested long ago by Albert Einstein (1950): any forms of matter or energy can be observed through their gravitational action, independent of their specific properties.

Let us finally write formulae which pertain to the future of a model of the Universe filled with neutrinos or photons (although this case is not characteristic of the actual Universe).

For $\Omega_0 < 1$, unbounded expansion will occur. For $\Omega_0 > 1$, the expansion will change to contraction after a time (see Fig. 5)

$$t_m - t_0 = \frac{1}{H_0}\left(\frac{1}{\Omega_0 - 1}\right), \qquad (1.7.2)$$

and the contraction will end upon the achievement of infinite density after the time

$$t_\infty - t_0 = \frac{1}{H_0}\left(\frac{1}{\Omega_0^{1/2}} - 1\right). \tag{1.7.3}$$

1.8 The Early Stages

Taking account of the law for the adiabatic expansion of neutrinos and photons, $\epsilon \sim l^{-4/3}$ or $\epsilon = \kappa/R^4$ with κ a constant, we obtain the equation for the change in radius of a sphere containing a fixed number of conserved particles:

$$\frac{d^2R}{dt^2} = -\frac{G}{R^2}\left(\frac{4\pi}{3}R^3\right)\left(\frac{1}{c^2}\frac{2\kappa}{R^4}\right). \tag{1.8.1}$$

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 - \frac{G}{R}\frac{4\pi}{3}R^3\frac{1}{c^2}\frac{\kappa}{R^4} = \text{const}. \tag{1.8.2}$$

At the beginning of the evolution near the time ($t = 0$) when the density is infinite, $R = 0$; hence, one can neglect the constant in the second equation. The solution then takes the form

$$\begin{aligned} R &= \left(\frac{32\pi G\kappa}{3c^2}\right)^{1/4} t^{1/2}; \qquad \epsilon = \frac{3c^2}{32\pi G t^2}; \\ \rho &= \frac{\epsilon}{c^2} = \frac{3}{32\pi G t^2} = \frac{4.5 \times 10^5}{t^2}\ \text{g cm}^{-3}, \end{aligned} \tag{1.8.3}$$

where t in the last expression is measured in seconds.

This expression for the density differs only by a numerical coefficient (~ 0.5) from the expression for the density in the case $P = 0$. It should be kept in mind, however, that the number density of conserved particles (nucleons) is

$$n = \text{const.}/R^3 = \text{const.}/t^{3/2}. \tag{1.8.4}$$

In distinction to the constant in equations (1.8.3), that in equation (1.8.4) is not determined by fundamental constants, but rather is a free parameter characterizing the evolution of the Universe. As t tends to zero, n grows more weakly than ρ, so the ratio $n/\rho \to 0$.

Notice that the number density of neutrinos and photons in this solution is also proportional to $R^{-3} \sim t^{-3/2}$. But the temperature and average energy of a single photon, and the average energy of a single neutrino, are all proportional to $t^{-1/2}$. From this it follows that the mass density, proportional to the product of the number of particles per unit volume and the energy per particle, is proportional to

$$t^{-3/2} \times t^{-1/2} = t^{-2}.$$

Let us finally examine the case of a cold nucleon gas under the assumption of the stiffest equation of state, i.e., with strong repulsion between the

nucleons. (See § 6.12 of Vol. I for more detail about the possibility of such an equation of state.) In this case,

$$P = \epsilon = \rho c^2 = \mathrm{B}n^2 , \tag{1.8.5}$$

where n is the density of nucleons. For an estimate of the constant, one can assume, depending upon the interaction, that the density is comparable to the rest mass density for ρ of order 100 ρ_{nuc}:

$$\rho = nm_0 = \frac{\mathrm{B}n^2}{c^2} = 10^{16}\ \mathrm{g\ cm^{-3}} . \tag{1.8.6}$$

Here, m_0 is the nucleon mass, so for $nm_0 = 10^{16}\ \mathrm{g\ cm^{-3}}$, it follows that $\mathrm{B} = 2.5 \times 10^{-43}$.

By the method that was used earlier, we find for this equation of state that

$$\rho = \frac{1}{24\pi G t^2} = \frac{2 \times 10^5}{t^2}\ \mathrm{g\ cm^{-3}} , \tag{1.8.7}$$

where t in the last expression is measured in seconds. Moreover, $R \sim t^{+1/3}$, so the nucleon number density satisfies

$$n \approx 2.7 \times 10^{34}/t\ \mathrm{cm^{-3}} . \tag{1.8.8}$$

Again, t in the above expression is in seconds.

By virtue of the condition $\rho > 100\ \rho_{\mathrm{nuc}}$, we see that this solution can only apply to a period not much longer than several microseconds.

Appendix to § 1.3

The reduction of the expansion time in comparison with the case of uniform expansion ($R \sim t$) depends on the density. The function $f(\Omega_0)$ characterizing this dependence is found via the solution of equation (1.2.4). We must use this formula for $P = 0$ since the overwhelming time of the evolution of the Universe is spent in the $P = 0$ domain. This function f satisfies $f \leq 1$. Moreover, for $\rho_0 = 0$, $\Omega_0 = 0$ and $f(0) = 1$. In the general case of arbitrary Ω_0, the function f is given by formulae which have different forms for $\Omega_0 > 1$ and for $\Omega_0 < 1$.

In the first case,

$$f = \frac{\Omega_0}{2(\Omega_0 - 1)^{3/2}}[\arcsin(2(\Omega_0 - 1)^{1/2}/\Omega_0) - 2(\Omega_0 - 1)^{1/2}/\Omega_0] . \qquad (1.3A.1)$$

Here, it is easy to verify that $f \to \frac{2}{3}$ for $\Omega_0 \to 1$. For $\Omega_0 \gg 1$, $f \approx \pi/2\Omega_0^{1/2}$, so that $f \to 0$ for $\Omega_0 \to \infty$. For the low-density case, $\Omega_0 < 1$,

$$f = \frac{\Omega_0}{2(1 - \Omega_0)^{3/2}}[2(1 - \Omega_0)^{1/2}/\Omega_0 - \operatorname{arcsinh}(2(1 - \Omega_0)^{1/2}/\Omega_0)] . \qquad (1.3A.2)$$

This formula, naturally, also gives $f \to \frac{2}{3}$ in the limit, $\Omega_0 \to 1$, so that $f(1)$ does not depend on the direction by which we approach $\Omega_0 = 1$. The function f does not exhibit a discontinuity or a kink notwithstanding the different analytic forms (1.3A.1) and (1.3A.2). For $\Omega_0 \ll 1$, we have the asymptotic form (see eq. [1.4.8] below)

$$f = 1 - \frac{\Omega_0}{2}\ln\frac{1}{\Omega_0} ; \quad f \to 1 \text{ for } \Omega_0 \to 0 . \qquad (1.3A.3)$$

For the case $\Omega_0 > 1$, two periods of time are of interest: that required for the achievement of a state of minimum density, $t_m - t_0$, and that required for the achievement of complete collapse, $t_\infty - t_0$ (see Fig. 1). From the solution of equation (1.2.4), one can show that

$$t_m - t_0 = \frac{1}{H_0} f_m(\Omega_0) , \quad t_\infty - t_0 = \frac{1}{H_0} f_\infty(\Omega_0) , \qquad (1.3A.4)$$

where

$$f_m(\Omega_0) = \frac{\pi\Omega_0}{2(\Omega_0 - 1)^{3/2}} - f(\Omega_0) . \qquad (1.3A.5)$$

Moreover,

$$f_\infty(\Omega_0) = +f + 2f_m . \qquad (1.3A.6)$$

The function f_m is displayed in Figure 5. The density passes through a minimum and then increases toward infinity only in the case $\Omega_0 > 1$. Correspondingly, formulae (1.3A.5) and (1.3A.6) are defined only for $\Omega_0 > 1$. For Ω_0 approaching unity, f_m and f_∞ diverge as $(\Omega_0 - 1)^{-3/2}$.

2 RELATIVISTIC THEORY OF THE HOMOGENEOUS, ISOTROPIC UNIVERSE

2.1 Einstein's Gravitational Equations and Friedmann's Cosmological Equations

In the previous chapter, the local properties of a homogeneous cosmological model were obtained from Newtonian theory. Of course, for a consistent picture of the cosmological problem, Newtonian theory is insufficient. Indeed, in § 1.1 we had to assume that in GTR, just as in classical Newtonian theory, no gravitational field is created inside an arbitrarily located spherical surface by the homogeneous, isotropically expanding matter outside. For proof, naturally, one must turn to GTR. In addition, the equations of GTR are absolutely necessary when we pass from the study of local properties to the study of large regions of space. In such regions, the self-gravity of the matter and the velocity of expansion are not small. An examination of such large regions is mandatory if we are to properly analyze the propagation of light from distant galaxies, galaxy counts, and so on.

Thus, having given in chapter 1 an elementary explanation of the essence of the expansion laws, let us now turn to a consistent deduction of them from the Einstein equations. As was noted in the Introduction, we assume that the reader is familiar with the fundamentals of GTR. Note though that for an understanding of the subsequent development, the information about GTR in Volume 1 is sufficient. If a complete account of GTR is necessary, that found in the last chapters of *The Classical Theory of Fields* by Landau and Lifshitz (1973) is unsurpassed with respect to clarity and economy of exposition.

We will construct a homogeneous, isotropic model of the Universe. In such a model, all points are equivalent and all directions are equivalent; nothing is chosen specially. The three-dimensional space of this model must be homogeneous and isotropic. To be sure, this three-dimensional space need not be a Euclidean space, since, in agreement with GTR, the geometrical properties of space depend on the density and the motion of matter. Any homogeneous, isotropic, three-dimensional space will do. From mathematics it is known that such spaces are spaces of constant curvature, (i.e., with curvature that is independent of direction and of the spatial coordinates). The square of the line element in such a space may be written

$$dl^2 = \frac{dx^2 + dy^2 + dz^2}{[1 + K(x^2 + y^2 + z^2)/4a^2]^2}, \qquad (2.1.1)$$

where $K = 0, \pm 1$. The quantity a has the dimensions of length. If $K = 0$, we are dealing with ordinary Euclidean space (sometimes called "flat" space) and x, y, z are ordinary Cartesian coordinates. If $K = 1$, we have a space of constant positive curvature, while if $K = -1$, we have a space of constant negative curvature. The quantity a is called the radius of curvature of the space, and the quantity $C_G \equiv K/a^2$, the curvature of the space. More precisely, $C_G^{-1/2}$ is called the radius of curvature of a space. Hence if $K = -1$, the radius of curvature is imaginary. We will, however, adhere to the terminology presented in the text and accepted in cosmology. We will postpone a consideration of the geometrical properties of these spaces until the following sections.

The radius of curvature a is the natural unit of length for the case $K \neq 0$. Therefore, let us change to the new dimensionless coordinates

$$\bar{x} = \frac{x}{a}, \quad \bar{y} = \frac{y}{a}, \quad \bar{z} = \frac{z}{a}. \qquad (2.1.2)$$

The squared line element (2.1.1) then becomes

$$dl^2 = a^2 \frac{(d\bar{x}^2 + d\bar{y}^2 + d\bar{z}^2)}{[1 + K(\bar{x}^2 + \bar{y}^2 + \bar{z}^2)/4]^2}. \qquad (2.1.3)$$

Expression (2.1.3) is also formally valid for $K = 0$, though in this case a is an arbitrary scale factor. For Euclidean space, the scale factor a is indeterminate and so may be chosen arbitrarily.

Our first problem is to write the expression for the four-dimensional interval $ds^2 = g_{ik}dx^i dx^k$. In a frame of reference in which the three-dimensional space is homogeneous and isotropic and for which the spatial metric takes the form (2.1.3), the matter must be at rest. (Indeed, if this were not the case, the velocity would select out a preferred direction at each point and consequently would destroy the isotropy.) Thus, the frame of reference (2.1.3) is comoving. Taking as the time coordinate the proper time of each particle, we then obtain $g_{00} \equiv c^2$. Finally, all components $g_{0\alpha}$ in ds^2 must vanish identically. (Latin indices run through the values 0, 1, 2, and 3; the index 0 refers to the time coordinate; while Greek indices run through the values 1, 2, and 3.) If they did not vanish, they would constitute a three-dimensional vector that would also destroy the isotropy. Hence, the expression for ds^2 can be written in the form

$$ds^2 = c^2 dt^2 - a^2 \frac{(d\bar{x}^2 + d\bar{y}^2 + d\bar{z}^2)}{[1 + K(\bar{x}^2 + \bar{y}^2 + \bar{z}^2)/4]^2}. \qquad (2.1.4)$$

In expression (2.1.4), the coordinates $\bar{x}$, $\bar{y}$, $\bar{z}$ are the Lagrangian coordinates of the particles of matter. This follows because the matter does not move

with respect to this frame of reference. The motion of the matter is described by the deformation of the reference frame itself. The scale factor in equation (2.1.4) depends only on the time: $a = a(t)$. Its growth with time describes the expansion of the reference frame, the increase of all distances within the reference system, and thus the expansion of the matter. Remember that the physical distance between any nearby particles in our model is dl and, in agreement with the line element (2.1.3), dl is proportional to $a(t)$ inasmuch as $d\bar{x}$, $d\bar{y}$, $d\bar{z}$ are constant. It remains for us to find the single unknown function $a(t)$ from the Einstein equations.

The Einstein equations are written in the following form:

$$R_i^k - \tfrac{1}{2}R\delta_i^k = -\frac{8\pi G}{c^4}T_i^k . \tag{2.1.5}$$

Here R_{ik} is the Ricci tensor, and R is its trace. They are functions of g_{ik} and its first and second derivatives. T_{ik} is the stress-energy tensor of matter. Here, where we shall take the matter to be a gas, the components of the stress-energy tensor are written in the form

$$T_i^k = (\rho c^2 + P)u_i u^k - P\delta_i^k , \tag{2.1.6}$$

where u^i is the four-velocity of the matter. In a comoving reference frame, (as in our case [2.1.4]), the nonzero components are

$$T_0^0 = \rho c^2 ; \quad T_1^1 = T_2^2 = T_3^3 = -P . \tag{2.1.7}$$

In a homogeneous, isotropic Universe, ρ and P can depend on the time but not on the spatial coordinates. Our task is to substitute the components g_{ik} from the line element (2.1.4) and the matter variables (2.1.7) into the Einstein equations (2.1.5). We can then determine the three unknown functions of time $a(t)$, $\rho(t)$, and $P(t)$. In reality, there are only two unknown functions, a and ρ, since ρ and P are connected by an equation of state.

Carrying out the above procedure, we are led after some simple transformations to the two equations[1]

$$\frac{d^2a}{dt^2} = -\frac{4\pi G}{3}a\left(\rho + \frac{3P}{c^2}\right); \tag{2.1.8}$$

$$\frac{1}{2}\left(\frac{da}{dt}\right)^2 - \frac{4\pi G}{3}\rho a^2 = -\frac{Kc^2}{2} . \quad K = 0, \pm 1 . \tag{2.1.9}$$

1. Direct substitution gives

$$\frac{8\pi G}{c^2}T_0^0 = \frac{8\pi G}{c^2}\rho c^2 = \frac{3}{a^2}\left(\frac{da}{dt}\right)^2 + \frac{3Kc^2}{a^2};$$

$$\frac{8\pi G}{c^2}T_1^1 = \frac{8\pi G}{c^2}T_2^2 = \frac{8\pi G}{c^2}T_3^3 = \frac{-8\pi G}{c^2}P = \frac{2}{a}\frac{d^2a}{dt} + \frac{1}{a^2}\left(\frac{da}{dt}\right)^2 + \frac{Kc^2}{a^2} .$$

These equations are connected by the identity

$$\frac{d\rho}{dt} + 3\frac{da/dt}{a}\left(\rho + \frac{P}{c^2}\right) = 0 . \tag{2.1.10}$$

The equations (2.1.8) and (2.1.9) describe the dependence of a on t (i.e., the dependence of all length scales in the model on time) and that of ρ on t.

Equations (2.1.8–10) are analogous to equations (1.6.2), (1.6.4), and (1.6.5), respectively. The only differences are that R is replaced by a and that the arbitrary constant in equation (1.6.5) is replaced by Kc^2 in equation (2.1.9). Recall that, in the previous chapter, the radius of the sphere was selected arbitrarily and that we were only interested in the dependence of the length scales on time. The equations of GTR give exactly the same results (for this time dependence) as did the Newtonian analysis.

Equations (2.1.8–10) are called the Friedmann cosmological equations. Their solutions were analyzed in detail in the previous chapter.

In conclusion, we present in Table 1 analytic solutions of these equations for the two important equations of state $P = 0$ and $P = \epsilon/3$. The solutions are written in parametric form.

2.2 Global Geometrical Structure of a Model of the Universe: The Space of Constant Positive Curvature

Let us now proceed to a consideration of the geometrical properties of the three-dimensional spatial sections of a homogeneous model. These properties are described by expression (2.1.3), which gives the three-dimensional spatial interval. It is clear that we are describing the geometrical properties of a homogeneous reference system at a fixed instant of proper time, or, as is sometimes said, the properties of a comoving space. In GTR one can select systems of reference arbitrarily. In our problem, one could also choose some other kind of system in place of the comoving system. In it the geometrical properties of the three-dimensional space would be different. Moreover, the matter would move relative to this system, and the properties of homogeneity and isotropy would not be satisfied in it. But what interests us primarily are the properties of the expanding matter in the Universe. This is why a comoving system is advantageous and why we primarily occupy ourselves with a study of the properties of three-dimensional space as described by such a system.

The properties of this space depend on one quantity—the curvature. This curvature, sometimes called the Gaussian curvature, is (see § 2.1)

$$C_G = Ka^{-2} , \quad K = \pm 1 , 0 , \tag{2.2.1}$$

TABLE 1

RADIUS OF CURVATURE, HUBBLE PARAMETER, AND DENSITY PARAMETER AS FUNCTIONS OF TIME

Pressure P	K	Time t	Radius of Curvature a	Hubble Parameter H	Density Parameter Ω
0	1	$c^{-1}A(\eta-\sin\eta)$	$A(1-\cos\eta)$	$A^{-1}c\sin\eta(1-\cos\eta)^{-2}$	$2(1+\cos\eta)^{-1}$
	0	$c^{-1}A\eta^3/6$	$A\eta^2/2$	$A^{-1}4c\eta^{-3}=\frac{2}{3}t^{-1}$	1
	-1	$c^{-1}A(\sinh\eta-\eta)$	$A(\cosh\eta-1)$	$A^{-1}c\sinh\eta(\cosh\eta-1)^{-2}$	$2(\cosh\eta+1)^{-1}$
$\epsilon/3$	1	$c^{-1}A'(1-\cos\eta)$	$A'\sin\eta$	$A'^{-1}c\cos\eta(\sin\eta)^{-2}$	$(\cos\eta)^{-2}$
	0	$c^{-1}A'\eta^2/2$	$A'\eta$	$A'^{-1}c\,\eta^{-2}=\frac{1}{2}t^{-1}$	1
	-1	$c^{-1}A'(\cosh\eta-1)$	$A'\sinh\eta$	$A'^{-1}c\cosh\eta(\sinh\eta)^{-2}$	$(\cosh\eta)^{-2}$

NOTES.—(1) A and A' are arbitrary constants; (2) c is the velocity of light; (3) $d\eta = cdt/a$; (4) $\Omega = 1+Kc^2/a^2H^2$; $\Omega(t) \equiv \rho(t)/\rho_c(t)$, where $\rho_c(t) = 3H^2(t)/8\pi G$ and $H(t) = a^{-1}\,da/dt$; (5) when $K = 0$, the radius of curvature is infinite; the above entry is for the scale factor.

here K and a are the quantities entering into formulae (2.1.3) and (2.1.9). The quantity $(|C_G|)^{-1/2} = a$ is called the radius of curvature. Substituting into definition (2.2.1) the quantity Ka^{-2} from equation (2.1.9), we obtain

$$C_G = \frac{8\pi G\rho}{3c^2} - \frac{1}{c^2}\left(\frac{1}{a}\frac{da}{dt}\right)^2. \qquad (2.2.2)$$

Thus, the geometrical properties of the space depend on the density and motion of the matter present. In equation (2.2.2) it is apparent that the sign of the curvature is determined by the same criterion regarding the density which appeared in § 1.2. If $\rho_0 > \rho_c$ (see eq. [1.2.5]), the curvature is positive: $C_G > 0$. If $\rho_0 < \rho_c$, the curvature is negative: $C_G < 0$.

The magnitude and sign of the curvature are the same at all points at a given moment. In addition, since either one of the inequalities $\rho(t) > \rho_c(t)$ or $\rho(t) < \rho_c(t)$ is preserved for all times, so also is the sign of the curvature.

From differential geometry it is known that a three-dimensional space of constant negative curvature extends without bound. Consequently, if the density of matter is less than the critical density, $\rho_0 < \rho_c$, then the homogeneous Universe is infinite.[2] This means that the model is topologically similar to an infinite Euclidean space. Such a model is called the "open model" of the Universe.

If $\rho_0 > \rho_c$, however, then the curvature is positive, so the three-dimensional space is closed and finite but has no boundary. Since our intuition is based on everyday experience and relates to three-dimensional Euclidean space, it is impossible to visualize a closed Universe (or, by the way, an open Universe). One can merely study its properties mathematically, compare the results of calculations with experiments and observations, and explain these results with the help of analogies and models, however incomplete. To study the properties of a closed Universe, let us proceed by taking one point to be the origin of coordinates and construct around it a sphere (i.e., consider the set of particles equidistant from the origin). Then, keeping in mind that this model of the Universe is not stationary, we shall determine such quantities as the equatorial length and surface area of the sphere. Since the length of the equator and surface area of the sphere formed by the given set of particles depend on the time of measurement, we shall agree to measure all quantities at a moment when the proper time of all the particles is the same. The proper time is reckoned from the moment of infinite density according to clocks which move with the particles. Here we are considering only the motion of particles corresponding to the general expansion. We are not taking account of the deviation in the velocity of

2. In this chapter and in subsequent chapters we will examine the simplest models from the topological point of view. Note though that the appearance of the concept of spacetime curvature in GTR allows one to consider the possibility of more complicated topological models. We will analyze this problem in §§ 23.12, 23.13.

individual galaxies from that velocity corresponding to a strictly homogeneous expanding Universe.

The most important conclusion of the theory is that in the case of a closed world ($\rho_0 > \rho_c$), as we consider more and more distant spheres, the length of the equator and the area of the surface of the sphere at first grow, then pass through a maximum, and finally decrease to zero. The meanings of "nearer" and "farther" are completely unambiguous—as are those of "inside" and "outside": a more distant sphere contains not only all that matter which is located within a nearby sphere, but also that matter located between the spheres.

A more distant sphere, containing more matter and having a greater volume, can at the same time have a shorter equator and a smaller surface. This is unusual and not similar to the situation in Euclidean geometry. However, such an uncommon situation is not logically contradictory.

Well-known analogous properties exist in a closed, curved two-dimensional space—the surface of an ordinary three-dimensional sphere. Take the north pole as a center. On the surface of the sphere, circles (lines of latitude) are analogous to the spheres discussed above. The circumference of such a circle at first grows in proportion to the distance from the north pole, but reaches a maximum at the equator. The circumference of more distant circles decreases. Meanwhile, the area bounded by a circle grows monotonically. Finally, as we approach the south pole, the area contained within the circles becomes equal to $4\pi r^2$, where r is the radius of the sphere. The circumference of the circles at the same time tends toward zero. Note that in the case of a closed world, a sphere separates all of space into two parts, each of which is finite. By the volume contained within a sphere, we always mean that part of the space which includes the origin of coordinates. In the two-dimensional analogy, we mean by the surface bounded by a line of latitude the surface of the part of the sphere lying north of the line of latitude and including the north pole. According to such a definition, the accumulation of the circles of latitude at the south pole is accompanied by the approach to $4\pi r^2$ of the area of the surface bounded by the latitudes.

Let us return to the closed three-dimensional space, which represents a homogeneous Universe when $\rho_0 > \rho_c$. Consider a sphere constructed around the origin of coordinates. The approach to zero of the equatorial length and surface area of this sphere, given a sufficiently large increase in its radius, just signifies that the three-dimensional space being considered is closed. We will now write the formulae corresponding to this situation.

Let us first rewrite the expression for the square of the line element (2.1.3) using different coordinates. Transforming to spherical coordinates

$$\bar{x} = \bar{r} \sin\theta \cos\phi\,, \quad \bar{y} = \bar{r} \sin\theta \sin\phi\,, \quad \bar{z} = \bar{r} \cos\theta\,, \qquad (2.2.3)$$

obtain the squared line element in the form

$$dl^2 = a^2\left[\frac{dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)}{(1 + Kr^2/4)^2}\right], \quad K = 0, \pm 1. \qquad (2.2.4)$$

For the case of a space of constant positive curvature, $K = 1$, we make the following replacement of the radial coordinate:

$$\frac{\bar{r}}{1 + \bar{r}^2/4} \equiv \sin r. \qquad (2.2.5)$$

As a result we find

$$dl^2 = a^2[dr^2 + \sin^2 r(d\theta^2 + \sin^2\theta d\phi^2)]. \qquad (2.2.6)$$

Here a is the radius of curvature of the three-dimensional space. Of course, transformation of the spatial coordinates will only change the expression for the squared length of the spatial separation. As before, the expression for ds^2 has the form $ds^2 = c^2dt^2 - dl^2$, where dl^2 can be written in any system of spatial coordinates, such as the one used above.

From the squared line element (2.2.6) it follows that the area of a sphere with radius r and centered on the origin is

$$S(r) = 4\pi a^2 \sin^2 r. \qquad (2.2.7)$$

The circumference of the equator of such a sphere is

$$l_{eq} = 2\pi a \sin r. \qquad (2.2.8)$$

From (2.2.7) it is apparent that S achieves a maximum $S_{max} = 4\pi a^2$ when $r = \pi/2$ and then decreases, reaching zero when $r = \pi$. The volume contained within the sphere of radius r is

$$V = \int_0^r S(r')a\,dr' = 4\pi a^3 \int_0^r \sin^2 r'\,dr' = 4\pi a^3\left[\frac{r}{2} - \frac{1}{4}\sin 2r\right]. \qquad (2.2.9)$$

Naturally, the volume grows monotonically with increasing r. For $r = \pi$ we obtain the full volume of the closed Universe:

$$V_{closed} = 2\pi^2 a^3. \qquad (2.2.10)$$

Correspondingly, knowing the average number density of nucleons n, one can find the total number of nucleons in the entire closed model: $N_{tot} = nV_{closed}$.

With the help of formula (2.2.2), let us write the expression for the radius of curvature a:

$$a = \left(\frac{1}{C_G}\right)^{1/2} = \frac{c}{H}\frac{1}{(\Omega - 1)^{1/2}}. \qquad (2.2.11)$$

In this section, the notation $\Omega = \rho(t)/\rho_c(t) = 8\pi G\rho(t)/3H^2$ is used. Then expression (2.2.11) treats only the case $\Omega > 1$. For Ω near unity, the radius

is very large and tends to infinity as the density tends to the critical value for closure, i.e., as $\Omega \to 1$.

Consider ordinary matter with density $\rho_0 = nm_0$, where m_0 is the mass of a nucleon and n the number density of nucleons. For a fixed value of H it turns out that the total number of nucleons in the Universe is smaller, the greater their density! Indeed for $n \gg \rho_c/m_0$, we have that $\Omega \gg 1$, so that $a \sim n^{-1/2}$. Since the volume varies as $V \sim a^3 \sim n^{-3/2}$, the total number satisfies $N_{\text{tot}} = nV \sim n^{-1/2}$. With the decrease of n the total number N grows, at first as $n^{-1/2}$, and then faster, since, with the approach of n to $n_c = \rho_c/m_0$, $N \to \infty$.

Now recall that the model we are examining is not static. The time dependence of the distance $R = R(t)$ between any pair of points was studied in the preceding sections. It is obvious that the dependence of the radius of a curvature a on time is exactly the same as for R since $R = \text{const.} \times a$. Thus both functions grow, achieve a maximum, and then fall to zero. This time dependence is shown in Figure 1.

2.3 The Metric of the Open Model

In the previous section, the properties of the closed model were examined. The Universe is closed if it is homogeneous (so that, in particular, its density and the Hubble parameter are both uniform in space at a given moment of time) and if the density condition $\rho_0 > \rho_c = 3H_0^2/8\pi G$ holds. However, recent observational data suggest that in reality, the opposite inequality, ρ_0/ρ_c, holds. In this case, as is apparent from equation (2.2.2), the curvature is negative, and we have the three-dimensional Lobachevsky space of constant negative curvature. Let us write the expression for dl^2 in this case. We first transform the squared line element, by means of the replacement of the radial coordinate

$$\frac{\bar{r}}{1 - \bar{r}^2/4} = \sinh r \,. \tag{2.3.1}$$

We then obtain

$$dl^2 = a^2[dr^2 + \sinh^2 r(d\theta^2 + \sin^2 \theta d\phi^2)] \,. \tag{2.3.2}$$

The three-dimensional Lobachevsky space is infinite (with respect to volume). In this case, the model is termed "open" (i.e., its qualitative topology does not differ from that of ordinary, three-dimensional Euclidean space). In particular, assuming as earlier that the Universe is homogeneous, we draw the conclusion that the total number of galaxies, stars, and nucleons in the Universe is in this case infinite.

Nevertheless, when $\rho_0 < \rho_c$ we have a metric significantly different from that of Euclidean space. We have a metric such that the quantitative geometrical properties of physical space differ from those of Euclidean space. From (2.2.2) it follows that the curvature C_G is negative. In the

comoving space, therefore, the sum of the angles of a triangle is less than π. Moreover, the length of the equator and the area of a sphere are greater than the corresponding Euclidean quantities at a given distance from the origin.

In the next section, we will discuss the limiting case of the open model, namely when $\rho_0 \to 0$.

7.1 The Limiting Case of Small Matter Density

The limiting case of small matter density, when $\rho_0 \ll 3H_0^2/8\pi G$, is distinctive in a particularly simple way. To the degree that one can consider $\rho_0 = 0$, one can neglect the forces of gravity. As was already noted, there is necessarily a period in the past of such a model when the force of gravity is not negligible (see § 1.4). However, we will not be examining this period.

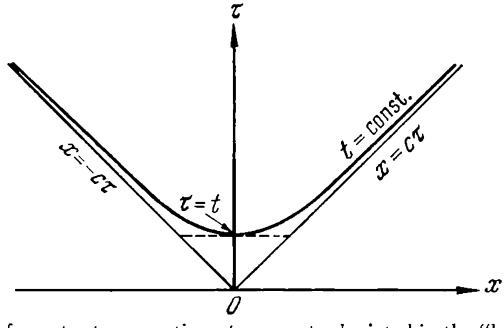


Fig. 6. A line of constant proper time, t = const., depicted in the "laboratory" system of coordinates in the Milne model. The dashed line τ = const. is a line of constant laboratory time.

So let us return to the period when $\rho(t) \ll \rho_c(t)$ and therefore when the forces of gravity are negligibly small.

One can picture the motion here quite readily in the following way. Suppose that in flat Euclidean space at an initial time $t = 0$ a set of particles moving with all possible velocities $\boldsymbol{u}$ flies out from the origin of coordinates. Of course, these velocities do not exceed the speed of light. We neglect the masses of the particles, so for consistency we also neglect the curvature of space which these particles could generate. Consequently, the motion occurs in flat (Euclidean) space, or, to be precise, in Minkowskian spacetime. The transformation from the metric of flat space to that comoving with the particles is described in § 1.4 of Volume 1. The comoving space of this system, as one can see from (2.2.2), has negative curvature.

In Figure 6 a section of the space in the "laboratory" coordinates τ and x is shown. In these coordinates, the velocities of the particles are less than c, so that all the trajectories are straight and lie within the light cone (i.e., within the angle formed by the straight lines $x = \pm c\tau$). Thus, $x = u\tau$,

with $u < c$, and the length of the circle drawn through the point x with center at $x = 0$ equals $2\pi x$. The proper time t of the particle moving with velocity u is expressed by means of the well-known formula from the theory of special relativity:

$$t = \tau(1 - u^2/c^2)^{1/2} \tag{2.4.1}$$

The set of points at which the particles are located after the same proper time $t = \text{const.}$ (see Fig. 6) is determined by the expressions

$$\tau = \frac{t}{(1 - u^2/c^2)^{1/2}}; \qquad x = u\tau = t\frac{u}{(1 - u^2/c^2)^{1/2}}. \tag{2.4.2}$$

These expressions constitute the equation of a hyperbola. Indeed, in order to construct the line $t = \text{const.}$ in the coordinates of Figure 6, we must specify t and τ, find u, and then, knowing u, express x via t and τ. We obtain

$$u = c\left[1 - \left(\frac{t}{\tau}\right)^2\right]^{1/2}; \qquad x = \frac{c}{[1 - (t/\tau)^2]^{1/2}}. \tag{2.4.3}$$

For fixed values of t, the curve giving the dependence of x on τ is a hyperbola with the lines $x = \pm c\tau$ as asymptotes.

Now, if the distance covered per unit of proper time is called the "velocity" u', then

$$u' \equiv \frac{x}{t} = \frac{u}{(1 - u^2/c^2)^{1/2}}, \tag{2.4.4}$$

and we obtain a quantity which can be larger than c. So a paradox arises that is connected with the question of whether there can be particles moving faster than the speed of light in the Friedmann solution. The resolution is that such velocities are obtained through the use of the proper time of a particle and of distances measured in the laboratory frame of reference. In other words, this is not the physical velocity of a particle.

For the calculation of a real velocity, it is necessary to measure both the distance and the time in the same system of reference. Then $u < c$ always holds. Laboratory experiments show that the meson (pion), with a lifetime of 2×10^{-8} s, can traverse a path significantly greater than $2 \times 10^{-8}\,\text{s} \times 3 \times 10^{10}\,\text{cm s}^{-1} = 6 \times 10^2$ cm if its velocity is near the velocity of light.

In determining the three-dimensional space from the condition of the equality of the proper time t in the Friedmann solution (in the comoving space), we obtained a space differing from the three-dimensional "laboratory" space determined by the condition $\tau = \text{const.}$ For a limiting density $\rho \to 0$, the entire four-dimensional spacetime is flat; in this limit its four-dimensional spacetime curvature vanishes. The world is Euclidean, or, more precisely, pseudo-Euclidean (Minkowskian), by virtue of the special role

played by time: $s^2 = r^2 - x^2 - y^2 - z^2$. But a three-dimensional section $\tau = \text{const}$ is not flat.

Consider a constant-laboratory-time spatial section of the set of world lines within the light cone $x = \pm c\tau$ on Figure 6, a section represented by the dashed line. Zel'manov (1959*a*) has emphasized that such a section is finite. Indeed, the spatial coordinate x is limited by $-c\tau \leq x \leq c\tau$. From this follows the conclusion that assertions about the finite or infinite character of a space are not invariant, but depend on how one chooses "space" and "time." Here we have considered only the simplest example; for the general case see Hawking and Ellis (1973).

Let us now examine the real case when the density ρ_0, although small, is nonzero. Particles whose proper time is small (see eq. [2.4.1]) are located in a section $\tau = \text{const.}$ near the boundary $x = c\tau$. Therefore, the density at the edge grows. One can say that in the section $\tau = \text{const.}$ the density equals $\rho_0[1 - (x/c\tau)^2]^{-1}$. Consequently, even for small ρ_0, there is in the section $\tau = \text{const.}$ a region where one may not neglect the density or the curvature caused by the gravitational field of the matter. For small but nonzero ρ_0, one cannot introduce a "laboratory time" over the whole space. The paradox about the finite character of the space thus disappears.

If one asks for the total number of conserved particles in the Universe, then from the very beginning there is no paradox at all because the quantity N is an invariant in GTR. Earlier it was stated that in a closed Universe the number N is finite. We even cited a concrete formula for N_{tot}. With $\rho_0 = \rho_c$ in an open world, N is infinite, a conclusion that does not depend on the method of calculation. In a section $t = \text{const.}$, the number density is constant, $n = n_0$, but the volume of a spatial section is infinite, so $N = \infty$. But even when one considers a section $\tau = \text{const.}$, when $\rho_0 \ll \rho_c$, $N = \int n_0 dV = \infty$ because the number density of particles at the boundary, $n = n_0[1 - (x/c\tau)^2]^{-1}$, grows without bound. The integration over the finite volume $V = (4\pi/3)(c\tau)^3$ thus gives an infinite value for N.

We conclude with a historical remark. The solution described above is called the Milne model. Indeed, Milne himself (1935, 1948) first introduced this set of particles which are not subject to gravity and which simultaneously emanate from one point with all velocities.

2.5 The Critical-Density Case

We will now discuss for a moment the case for which

$$\rho(t) = \rho_c(t) = \frac{3H^2}{8\pi G}. \tag{2.5.1}$$

If this relation holds at a definite instant (for example, at present), then the equation of motion guarantees that $\rho(t) = \rho_c(t)$ is satisfied for all times in the past as well as in the future (even when the pressure is nonzero).

When $\rho = \rho_c$, the metric of the world is especially simple—namely, in Friedmann coordinates the three-dimensional section $t = \text{const}$. (t being proper time) is Euclidean, and the line element (2.2.4) is written in the form

$$dl^2 = a^2[dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)]\,. \qquad (2.5.2)$$

This is the ordinary expression for the distance in Euclidean space in spherical coordinates; a is an arbitrary scale factor. For curved space, when $K = \pm 1$, the factor a is not arbitrary, but is equal to the radius of curvature of the space. This radius is naturally the scale for the measurement of all lengths. For the case of Euclidean space, this is not so. The radius of curvature is infinite, thereby setting no scale, so the expression (2.5.2) permits scale changes. The factor a is therefore arbitrary. When we turn to an expanding model with the critical density (2.5.1), then, in the expression for the interval

$$ds^2 = c^2dt^2 - a^2(t)[dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)]\,, \qquad (2.5.3)$$

the factor $a(t)$ is determined up to a numerical constant not dependent on time.

In the simplest case, with $P = 0$, we already know that for $\rho_0 = \rho_c$, the distance between every pair of points grows in proportion to $t^{2/3}$: $a = a_1 t^{2/3}$. The constant a_1 depends on the choice of scale. Indeed, one can always find a time t_1 such that $a_1 t_1^{2/3} = 1$ for any a_1.

Thus, the expression $l = a_1 t_1^{2/3} r$ at the time t_1 coincides with the ordinary distance if the radius is expressed in centimeters. Inasmuch as the time t_1 is arbitrary, the constant a_1 is arbitrary. Only after the choice of t_1 and a_1 is the scale of the coordinate r determined.

When $\rho_0 = \rho_c$, all calculations are simplified. As was already noted, the difference between the density and the critical density is always small at the beginning. Therefore, in the limit when the density is large, one can always carry out calculations using the relations of the flat (critical density) model.

How is one to understand the apparent flatness of the world for a nonzero matter density? In this way. The whole four-dimensional manifold of spacetime is by no means flat. When $\rho_0 = \rho_c$, only certain well-defined three-dimensional sections of spacetime are flat. In this case, these sections are the ones of equal proper time as registered by a mutually receding set of particles.

3 THE PROPAGATION OF PHOTONS AND NEUTRINOS; OBSERVATIONAL METHODS FOR TESTING COSMOLOGICAL THEORIES

3.1 The Redshift and Momentum Decrease

The redshift of light in an expanding Universe is analogous to the Doppler effect. Suppose that a photon emitted at the time t_1 from the point l_1 with a frequency ω_1 arrives at the nearby point l_2 at the time t_2. The change in its frequency is given by the formula

$$\omega_1 - \omega_2 = \omega_1 \left(\frac{u_{21}}{c} \right), \tag{3.1.1}$$

where u_{21} is the velocity of the matter at point 2 relative to the matter at point 1. The formula is valid for the case when the velocity is directed along the line joining the points 1 and 2, i.e., along the line of propagation of the photon. If l_{21} is small, u_{21} is small: $u_{21} \ll c$. For the Hubble expansion, the velocity satisfies $u_{21} = Hl_{21}$, so that

$$\omega_1 - \omega_2 = \frac{1}{c} H l_{21} \omega_1 .$$

This expression is asymptotically exact in the limit of small l_{21} since all corrections, namely those arising from the curvature of space and from relativistic effects not included in formula (3.1.1), are of order $(u/c)^2$, i.e., are of much higher order in l_{21}. Therefore, in differential form we have the exact equality

$$d\omega = -\frac{1}{c} H \omega dl . \tag{3.1.2}$$

Whether the photon with frequency ω_1 was emitted at the point l_1 or whether it arrived at this point from some other distant point and had the frequency ω_1 at l_1 is irrelevant. It is essential only that ω_1 be measured in the system moving together with the matter located at l_1, and that ω_2 be analogously measured at l_2.

One can consider the equality (3.1.2) as a differential equation which enables one to follow the changes in frequency of the photon step by step—in particular, in that case when the path traversed by the photon and its change in frequency are by no means small.

In the subsequent analysis we will use

$$\frac{1}{c}\,dl = dt \quad \text{and} \quad H = \frac{1}{a}\frac{da}{dt} = \frac{d(\ln a)}{dt}\,.$$

Here, a is the scale factor. However, one can equally well write

$$H = \frac{1}{l_{21}}\left(\frac{dl_{21}}{dt}\right),$$

where l_{21} is the distance between two fixed (i.e., comoving) particles and is proportional to $a(t)$. Consequently,

$$d\omega = -\omega\,\frac{d(\ln a)}{dt}\,dt\,,$$

so

$$\ln\omega + \ln a = \text{const.}\,; \qquad a\omega = \text{const.}\,; \qquad \omega = \omega_0(a_0/a)\,. \quad (3.1.3)$$

The frequency changes inversely in proportion to the scale factor.

In a completely analogous way one can examine the motion of a material particle. By its velocity w, we will mean the velocity relative to the average Hubble expansion at that point where the particle is instantaneously located. For simplicity, we consider nonrelativistic motion $w \ll c$. Then, by the velocity-addition law,

$$w_2 = w_1 - u_{21} = w_1 - Hl_{21}\,, \qquad w_2 - w_1 = dw = -Hdl\,.$$

We substitute $dl = wdt$ and $H = a^{-1}da/dt$ to obtain

$$dw = -\frac{1}{a}\frac{da}{dt}\,wdt\,,$$

so again,

$$w = \text{const.}/a\,. \quad (3.1.4)$$

In the nonrelativistic case, the momentum of a particle is $p = mw$ so that $p \sim 1/a$. For the general case, when w may be comparable to the speed of light, one can show, using the relativistic velocity-addition formula, that the law $p \sim 1/a$ continues to hold. Then, however, $p = mw/(1 - w^2/c^2)^{1/2}$ and the velocity w is not inversely proportional to a.

The redshift law for photons naturally agrees with the law of decreasing momentum. The momentum of a photon is $p = \hbar\omega/c$ while its wavelength is $\lambda = 2\pi c/\omega = 2\pi\hbar/p$. Indeed, the quantum-mechanical (de Broglie) wavelength of any particle can be expressed through p. Thus, both laws—that of the redshift of photons and that of the momentum change of an arbitrary particle—can be jointly formulated in the following way: In an expanding, homogeneous model of the Universe, all wavelengths change in proportion to the change of all distances, i.e., in proportion to the scale factor.

The following analogy is useful. Consider the case of a closed model, and suppose that a standing electromagnetic wave is located in it. Here we are thinking of the Universe as a hollow resonator in which a definite oscillation

definite harmonic) is excited—for example, that with n nodes. During the expansion the number of the harmonic and the number of nodes remains unchanged, it is therefore natural that the wavelength grows in proportion to the "radius" of the Universe. This interpretation has been repeatedly noted; see, for example, Wheeler (1958, 1960), von Laue (1931), Pauli (1958).

While considering the change in frequency for successive, small displacements dl along a path, we did not take account of the change in the frequency due to gravitation which is of order $(dl)^2$. Nevertheless, after integration we obtain an exact result for large displacements. It is important that the frequency be measured by observers moving like free particles, uninfluenced by pressure gradients, for only in the local reference frame of such observers will the gravitational acceleration vanish at the origin. In this situation, the gravitational potential, normalized to zero at the origin, is of order $(dl)^2$ at a distance dl from the origin, and such a term will not affect the differential law (3.1.2). On the other hand, for a direct calculation of the change in frequency along a finite path l_{21}, it is evident that one would have to take account of the gravitational effect on the frequency (see § 3.4).

The calculation of the redshift as an integral over the shift between neighboring free particles is an exact procedure. It applies not only for the homogeneous, isotropic Universe, but also to models where perturbations, inhomogeneity, and anisotropy are present.

3.2 Observable Quantities and the Horizon

A comparison of the predictions of the theory of the expanding Universe with observational data is difficult; unresolved problems remain. For the study of cosmological questions, we will be concerned with distances over which the curvature of space manifests itself. In particular, interest in large distances has grown with the discovery of large-redshift quasars. For such large distances one may not use Euclidean geometry with the usual notions about space that it entails. (The question of the limits of applicability of the Euclidean approximation will receive special attention in § 3.5.) For example, the concept of "distance" itself does not have a definite, unambiguous meaning. Hence, some preliminary mathematical work is necessary to derive relations among the observational quantities. This theoretical formulation is especially emphasized in the well-known reviews of McVittie (1959*a*, 1962*a*, *b*); see also the work of Sandage (1961) and of Mattig (1958, 1959).

Hubble's constant H_0 and the present value of the matter density, ρ_0, enter all formulae of the theory. Hubble's constant will be considered known to an accuracy of $\pm 50\%$ (see § 3.9). The density of matter is only known to lie within the rough limits 3×10^{-31} g cm^{-3} and 10^{-28} g cm^{-3} due

to the possible presence of large amounts of difficult-to-observe forms of matter. As a result, relations among the observational quantities for different assumed values of ρ_0 and H_0 are often presented. (Below, a summary of the corresponding formulae and curves will be presented.) One should also recall the possibility that $\Lambda \neq 0$. This case will be discussed in § 4.1 and § 4.2.

In principle, a comparison of observations with the formulae in which the density appears as a parameter must lead to a definite value for the density. However, the parameters of the various sources are insufficiently well known to achieve this goal. Moreover, we observe today light that was emitted long ago, i.e., emitted by objects at much earlier stages in their evolution than nearby objects. Therefore, interpretation of the observations depends on assumptions concerning unknown evolutionary effects. To this day, therefore, all conclusions remain very uncertain.

Let us begin with some qualitative features of the observable picture. There is a time when $\rho = \infty$, $a = 0$. We agree to denote this time by $t = 0$. It is clear that one cannot see light emitted earlier, because the optical depth, $\tau \approx \int \sigma n dx$, where σ is the photoabsorption cross section, diverges as $t \to 0$.

One can determine those particles which must emit photons at $t = 0$ so that those photons (propagating without absorption) are received by us today (at the time $t = t_0$). The set of such particles forms a sphere centered on the point of observation. This sphere is called the "particle horizon." In principle we can only receive information from matter located within the horizon. In practice, because of the scattering and absorption of light, we can only see regions such that the light reaching us today left those regions when the average density of matter in the Universe satisfied $\rho < \rho(\tau_\gamma = 1) \approx 10^{-20}\,\mathrm{g\,cm^{-3}}$ [or $\rho < \rho(\tau_\nu = 1) \approx 1\,\mathrm{g\,cm^{-3}}$ for neutrinos]. Practically speaking, however, the calculation of the propagation of light during the expansion yields the same result for the period from the time at which the optical depth reaches unity to the present as it does for the period from $t = 0$ to the present. Note that our results are for the case $\Lambda = 0$.

The propagation of light is governed by the condition $ds = 0$. We place the origin at the point of observation, so that the photons received move toward the origin. We employ an interval written in the form

$$ds^2 = c^2 dt^2 - a^2(t)[dr^2 + \Phi^2(r)(d\theta^2 + \sin^2\theta d\phi^2)]\,, \qquad (3.2.1)$$

where $\Phi(r) = \sin r$ for $\Omega_0 > 1$ and $\Phi(r) = \sinh r$ for $\Omega_0 < 1$. For $\Omega_0 = 1$, $\Phi(r) = r$ and we will denote the scale factor by $b(t)$ instead of $a(t)$. A light ray propagating along the radius ($d\theta = d\phi = 0$) is described by the equation

$$0 = ds^2 = c^2 dt^2 - a^2 dr^2\,; \qquad -c dt = \begin{cases} a(t)dr\,; & \Omega_0 \neq 1 \\ b(t)dr\,; & \Omega_0 = 1\,. \end{cases} \qquad (3.2.2)$$

The minus sign appears because we consider light rays approaching an observer located at the origin. The corresponding present-day values of $a(t)$ (independent of the equation of state!) can be chosen to be

$$a_0 \equiv a(t_0) = \frac{c}{H_0}\left(\frac{1}{\Omega_0 - 1}\right)^{1/2}, \quad \Omega_0 > 1 ;$$
$$a_0 \equiv a(t_0) = \frac{c}{H_0}\left(\frac{1}{1 - \Omega_0}\right)^{1/2}, \quad \Omega_0 < 1 . \tag{3.2.3}$$

In the critical density case, $\Omega_0 = 1$, $\Phi(r) = r$ and, instead of $a(t)$, we write $b(t)$ with the arbitrary normalization $b(t_0) \equiv b_0 = c/H_0$.

As functions of time, $a(t)$ and $b(t)$ will depend on the equation of state. The function $a(t)$ is easily found by integration of the equations in § 2.1 and was presented in Table 1. For convenience we repeat the results here. For $P = 0$, we have:

$$\Omega_0 > 1 ; \quad a = A(1 - \cos\eta) , \quad ct = A(\eta - \sin\eta) , \quad \Omega = \frac{2}{1 + \cos\eta} ;$$
$$\Omega_0 < 1 ; \quad a = A(\cosh\eta - 1) , \quad ct = A(\sinh\eta - \eta) , \quad \Omega = \frac{2}{1 + \cosh\eta} ; \tag{3.2.4}$$
$$\Omega_0 = 1 ; \quad b(t) = Bt^{2/3} .$$

For $P = \epsilon/3$:

$$\Omega_0 > 1 ; \quad a = A' \sin\eta , \quad ct = A'(1 - \cos\eta) , \quad \Omega = \frac{1}{\cos^2\eta} ;$$
$$\Omega_0 < 1 ; \quad a = A' \sinh\eta , \quad ct = A'(\cosh\eta - 1) , \quad \Omega = \frac{1}{\cosh^2\eta} ; \tag{3.2.5}$$
$$\Omega_0 = 1 ; \quad b(t) = B't^{1/2} .$$

Recall that the parameter η satisfies $d\eta = cdt/a$; as a result we see that η is the dimensionless coordinate r of a photon which leaves the origin at $t = 0$. Equivalently, a photon emitted at $t = 0$ from η reaches the origin at time t.

In equations (3.2.4) and (3.2.5), $\Omega = \Omega(t) = 8\pi G\rho(t)/3H^2(t)$. If we specify the present-day values ρ_0 and H_0, then we may determine Ω_0; from Ω_0 we can determine η_0 and, hence, A (or A'). Note that η_0 depends only on Ω_0 [and, of course, the equation of state $P = P(\rho)$].

Once η_0 and A (or A') are known, the age of the Universe t_0 may be

found from solution (3.2.4) or solution (3.2.5). In the critical density case ($\Omega_0 = 1$) we find:

$$P = 0\,, \quad H_0 t_0 = \tfrac{2}{3}\,; \quad P = \epsilon/3\,, \quad H_0 t_0 = \tfrac{1}{2}\,. \tag{3.2.6}$$

For $\Omega_0 \gtrless 1$, t_0 has been presented earlier as a function of H_0 and Ω_0. Recall that for $P = \epsilon/3$ we have the simple result $H_0 t_0 = (1 + \Omega_0^{1/2})^{-1}$. For t_0 when $P = 0$, see § 1.3.

The present value of the comoving coordinate of the horizon, r_0, is given by

$$r_0 = c \int_0^{t_0} \frac{dt}{a(t)} = \int_0^{\eta_0} d\eta = \eta_0\,. \tag{3.2.7}$$

Note that even though for $t \to 0$, $a^{-1} \to \infty$, the integral does not diverge. In particular, for the critical density case,

$$\begin{aligned} P = 0\,, \quad & r_0 = \frac{3ct_0}{b_0} = \frac{2c}{b_0 H_0}\,; \\ P = \epsilon/3\,, \quad & r_0 = \frac{2ct_0}{b_0} = \frac{c}{b_0 H_0}\,. \end{aligned} \tag{3.2.8}$$

From r_0 (or η_0) one can compute two distances:

$$\begin{aligned} R_0 &= a_0 r_0\,; \\ \bar{R}_0 &= a_0 \Phi(r_0)\,; \quad \Phi(r_0) = \begin{cases} \sin r_0\,, & \Omega_0 > 1 \\ \sinh r_0\,, & \Omega_0 < 1\,. \end{cases} \end{aligned} \tag{3.2.9}$$

The first quantity, R_0, is the present-day radius of the particle horizon; the circumference of the present-day particle horizon is $2\pi\bar{R}_0$. For the volume V_0 contained within the horizon as well as N_0, the total number of nucleons inside the horizon, we have

$$\begin{aligned} V_0 &= 4\pi a_0^3 \int_0^{r_0} \Phi^2(r)\,dr\,, \\ N_0 &= n_0 V_0\,, \end{aligned} \tag{3.2.10}$$

where n_0 is the present nucleon number density.

For the case $P = 0$, the mass within the horizon is $M_0 = \rho_0 V_0$, where ρ_0 is the present mass density. When $P \neq 0$, the formulae (3.2.10) still apply; however, the mass in a comoving volume element is not constant due to the work done by the pressure.

It is convenient to write the results in a general form by separating off the dimensional factors:

$$\bar{R}_0 = \frac{c}{H_0}\, s(\Omega_0)\,;$$

$$V_0 = \frac{4\pi}{3}\left(\frac{c}{H_0}\right)^3 g(\Omega_0)\,;$$

The function g for two equations of state ($P = 0$, g_1; and $P = \epsilon/3$, g_2) is presented graphically in Figure 7. For reference, we give the formulae below, however, we warn the reader that they are cumbersome and that calculations involving them are complicated.

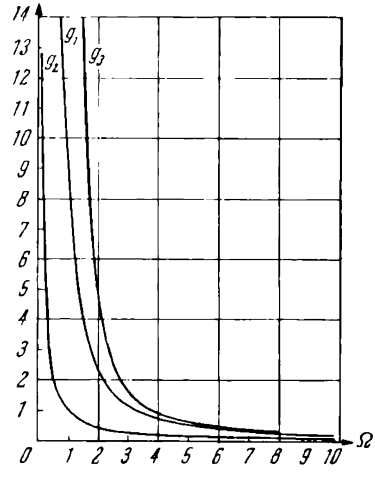


Fig. 7. The dimensionless volume g of the region of the Universe accessible to an observer (the region within the horizon) as a function of the density parameter Ω. Note that g is also the renormalized number of nucleons inside the horizon. The curve labeled g_1 is for the equation of state $P = 0$, while that labeled g_2 is for the equation of state $P = \epsilon/3$. For comparison purposes, the curve labeled g_3 is presented. It gives the total volume and the total number of baryons in a closed universe (defined only for $\Omega > 1$) for the equation of state $P = 0$.

For the equation of state $P = 0$, we find that the dimensionless functions $s(\Omega_0)$ and $g(\Omega_0)$ have these forms:

$$s = s_1(\Omega_0) = 2\Omega_0^{-1} ;$$

$$\Omega_0 < 1 , \quad g_1(\Omega_0) = \frac{3}{2(1-\Omega_0)^{3/2}} \left\{ \left(\frac{2}{\Omega_0} - 1\right) \left[\frac{4}{\Omega_0}\left(\frac{1}{\Omega_0} - 1\right)\right]^{1/2} - \operatorname{arc\,sinh} \left[\frac{4}{\Omega_0}\left(\frac{1}{\Omega_0} - 1\right)\right]^{1/2} \right\} ;$$

$$\Omega_0 = 1 , \quad g_1(\Omega_0) = 8 ; \tag{3.2.11}$$

$$\Omega_0 > 1 , \quad g_1(\Omega_0) = \frac{3/2}{(\Omega_0 - 1)^{3/2}} \left\{ \arccos\left(\frac{2}{\Omega_0} - 1\right) - \left(\frac{2}{\Omega_0} - 1\right) \left[\frac{4}{\Omega_0}\left(1 - \frac{1}{\Omega_0}\right)\right]^{1/2} \right\} .$$

In the case $P = \epsilon/3$, we obtain

$$s = s_2 = \Omega_0^{-1/2} ;$$

$$\Omega_0 < 1 , \quad g_2(\Omega_0) = \frac{3}{2(1-\Omega_0)^{3/2}} \left[\frac{(1-\Omega_0)^{1/2}}{\Omega_0} - \operatorname{arc\,sinh} \left(\frac{1}{\Omega_0} - 1 \right)^{1/2} \right] ; \tag{3.2.12}$$

$$\Omega_0 = 1 , \quad g_2(\Omega_0) = 1 ;$$

$$\Omega_0 > 1 , \quad g_2(\Omega_0) = \frac{3}{2(\Omega_0 - 1)^{3/2}} \left[\arcsin \left(1 - \frac{1}{\Omega_0} \right)^{1/2} - \frac{1}{\Omega_0} (\Omega_0 - 1)^{1/2} \right] .$$

Finally, let us present the asymptotic forms:

$$\Omega_0 \gg 1 , \quad g_1(\Omega_0) = \frac{3\pi}{2} \Omega_0^{-3/2} , \quad g_2(\Omega_0) = \frac{3\pi}{4} \Omega_0^{-3/2} ; \tag{3.2.13}$$

$$\Omega_0 \ll 1 , \quad g_1(\Omega_0) = 6\Omega_0^{-2} , \quad g_2(\Omega_0) = \tfrac{3}{2} \Omega_0^{-1} .$$

Pay particular attention to the continuity in the curves at the critical density, i.e., for $\Omega_0 = 1$. At this point a qualitative change occurs—the homogeneous model of the Universe changes from the infinite (open) model for $\Omega_0 \leq 1$ to the finite (closed) model for $\Omega_0 > 1$.

For $\Omega_0 \leq 1$, the total volume V_{TOT}, mass M_{TOT}, and number of particles N_{TOT} are infinite, while for $\Omega_0 > 1$ these quantities are finite. However, in both cases the part of the Universe accessible to observation is finite, and in this sense the differences between the open and closed models are less than one might have expected.

The total number of particles in the closed Universe N_{TOT} (for $\Omega_0 > 1$) can be expressed in a form containing a dimensionless function:

$$N_{\text{TOT}} = \frac{4\pi}{3} \left(\frac{c}{H_0} \right)^3 n_0 g_3(\Omega_0) ; \tag{3.2.14}$$

$$g_3(\Omega_0) = \frac{3\pi}{2} (\Omega_0 - 1)^{-3/2} .$$

The function $g_3(\Omega_0)$ is shown in Figure 7. Notice that, for $\Omega_0 \gg 1$, $g_1 \to g_3$, $g_2 \to g_3/2$. Thus, if the density is so large that the Universe is closed, then, in principle, we can observe almost all the matter in the Universe for the case $P = 0$, but only half of the matter for the case $P = \epsilon/3$.

Is it possible to observe the same object from both sides in the closed model? Using the analogy with the surface of the Earth, we can formulate the question as follows: If someone is located at the north pole, can he

communicate with a Moscow radio station once by means of a signal traversing the shortest path and, additionally, by means of a signal that traverses the great circle route from Moscow, through the south pole, and finally to the observer at the north pole?

On the Earth, this is possible using long-wavelength radio signals, which hug the Earth's surface. In the Universe, this is not possible; not enough time will have elapsed from the moment $t = 0$, when $\rho = \infty$. A photon emitted at that moment or later will not be able to reach a present-day observer via the longer path.

But if $\Omega_0 > 1$ and $P = 0$, this would be just possible after several billion years. In this case the expansion changes to contraction, so that our descendants will be talking about the (Hubble) blueshift. Moreover, a photon emitted at $t = 0$, when $\rho = \infty$, would be received at $t = t_\infty$, when $\rho = \infty$.

We emphasize that this is a question of theoretical possibilities, limited not by the present-day level of experimental technique, but by the finite value of the velocity of light. To be sure, even neutrino astronomy will not expand the horizon. Note that the horizon itself gradually expands with the passage of time in proportion to the decrease of H_0 and the increase of t_0. However, a noticeable increase in R_0, H_0, and V_0 would require a time of the order of t_0—billions of years! Consequently, statements about the finite or infinite character of the Universe are an extrapolation of the known situation in our neighborhood to regions not only far away, but also unobservable at present.

Of course, the concept of an horizon arises only because, in the period subsequent to the moment when $\rho = \infty$, light emitted by distant objects does not have sufficient time to reach an observer. This concept is also connected with the extrapolation to $t \to 0$ of the dynamics of the homogeneous and isotropic model (for a discussion of horizons in more general models, see Part III).

3.3 Graphs and Formulae for Functions Which Determine Observable Quantities

To characterize the redshift, we shall introduce the quantity

$$\Delta = 1 - \frac{\omega_{\text{rec}}}{\omega_{\text{emit}}} = \frac{\omega_{\text{emit}} - \omega_{\text{rec}}}{\omega_{\text{emit}}}. \tag{3.3.1}$$

This quantity differs from the frequently employed redshift z which satisfies

$$z = \frac{\lambda_{\text{rec}} - \lambda_{\text{emit}}}{\lambda_{\text{emit}}}. \tag{3.3.2}$$

Here, ω is the frequency and λ the wavelength of light. It is obvious that

$$z = \frac{\Delta}{1 - \Delta}, \quad \Delta = \frac{z}{1 + z}. \tag{3.3.3}$$

Δ varies within the finite region from $\Delta = 0$ (near the point of observation) to $\Delta = 1$ (at the horizon). The limit $\Delta \to 1$ corresponds to the frequency of the observed light tending to zero; $\Delta = 1$ is the limit for possible observation. The variable z varies between zero and infinity, making the construction of graphs awkward. Because the use of z is so widespread, we give formulae with Δ and with z. The graphs are presented in terms of Δ; for quick conversion to z, we give the following table:

Δ.....	0	0.2	0.4	0.5	0.6	0.8	1
z.....	0	0.25	0.67	1.0	1.5	4.0	∞ .

Note especially these limits: for $\Delta \ll 1$, $z \approx \Delta$; while for $1 - \Delta \ll 1$ ($\Delta \approx 1$), $z \approx (1 - \Delta)^{-1}$.

Let us now consider the angle θ subtended by an object of linear size l. In Euclidean geometry,

$$\theta = l/R\,; \quad R = l/\theta\,, \tag{3.3.4}$$

where R is the distance to the object. In a curved Universe, we shall by definition call l/θ the angular diameter distance and shall denote it by $\bar{R}$. Then $2\pi\bar{R}$ is the circumference of the circle whose center is the point of observation and on which the object of consideration was located at the moment of emission. Comparing this with the expression for the interval, we find $\bar{R} = a(t_e)\Phi(r_e)$, where r_e is the comoving coordinate of the object (see formula [3.2.1]). The time of emission t_e is fixed by the condition that the light emitted at coordinate r_e at time t_e is received today (t_0) by the observer ($r = 0$). Near the horizon, $t_e \to 0$ and $a(t_e) \to 0$, so $\bar{R} \to 0$ and the angle θ grows without bound.

From this it follows that for a given l, the angle θ passes through a minimum. The presence of a minimum in θ and a maximum in $\bar{R}$ is a general property of an expanding Universe, valid for both open and closed models. The curves showing the dependence of $\bar{R}$ on θ for several cases were given by Wheeler (1958, 1960), but only for the closed model. A new approach to the physical interpretation of this minimum will be set forth in connection with the calculations of the angles in an inhomogeneous Universe (see §3.10).

For comparatively nearby objects with $\Delta \ll 1$, the geometry is Euclidean. Then, one can consider the Doppler effect in the nonrelativistic limit $d\omega \approx -(u/c)\omega$, since the change in distance during the light travel time is small. Thus, in the limit of small Δ,

$$u \approx c\frac{d\omega}{\omega} \approx c\Delta \approx H_0\bar{R}\,, \quad \bar{R} \approx \frac{c}{H_0}\Delta\,; \tag{3.3.5a}$$

$$u \ll c\,, \quad R \approx \frac{c}{H_0}z\,. \tag{3.3.5b}$$

Let us express the result for any distance with the help of the dimensionless angular diameter distance Ψ:

$$R = \frac{c}{H_0}\Psi_\Delta(\Delta)\,, \tag{3.3.6a}$$

$$R = \frac{c}{H_0}\Psi_z(z)\,. \tag{3.3.6b}$$

One can show (see the following section) that the functions depend not only on the quantity Δ (or z), but also on the ratio $\Omega_0 = \rho_0/\rho_c$ of the present-day density ρ_0 to the present-day critical density $\rho_c = 3H_0^2/8\pi G$. Moreover, the form of Ψ for $P = 0$ differs from that for $P = \epsilon/3$. The general properties of these functions are

$$\Psi_\Delta(\Delta \to 0, \Omega_0) = \Delta\,, \quad \Psi_\Delta(\Delta = 1, \Omega_0 \neq 0) = 0\,; \tag{3.3.7a}$$

$$\Psi_z(z \to 0, \Omega_0) = z\,, \quad \Psi_z(z \to \infty, \Omega_0 \neq 0) = 0\,. \tag{3.3.7b}$$

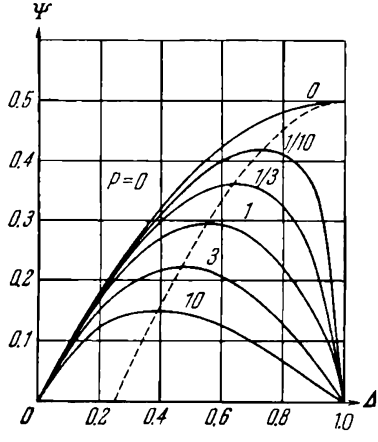


Fig. 8. The dimensionless angular-diameter distance Ψ (inversely proportional to the apparent angular diameter of an object with a fixed linear size) as a function of the redshift parameter Δ, for various values of the dimensionless density Ω and for the equation of state $P = 0$. The values of Ω are given by the numbers above the curves, and the dashed line traces the locus of the curve maxima.

In Figures 8 and 9 the curves $\Psi_\Delta(\Delta, \Omega_0)$ are presented for $\Omega_0 = 0$, 1/10, $\frac{1}{3}$, 1, 3, 10. Figure 8 contains the curves for $P = 0$ while Figure 9 contains those for $P = \epsilon/3$. The dashed curve represents the locus of the maxima of Ψ_Δ. Note carefully the limiting case $\Omega_0 = 0$: the smaller the value of Ω_0, the closer the maximum of the curve is to unity. In the limit $\Omega_0 = 0$, i.e., for the Milne model, the function takes the simple form

$$\Psi_\Delta(\Delta; 0) = \Delta\left(1 - \frac{\Delta}{2}\right); \tag{3.3.8a}$$

$$\Psi_z(z; 0) = \tfrac{1}{2}[1 - (1 + z)^{-2}]\,. \tag{3.3.8b}$$

The curves for $\Omega_0 = 0$ on the two graphs (for $P = 0$ and for $P = \epsilon/3$) naturally coincide. For $\Omega_0 = 0$ the function has no maximum, or, more precisely, the maximum is located at the edge of the domain of definition of the function, at $\Delta = 1$.

For high density, $\Omega_0 \gg 1$, the maximum of Ψ_Δ lies at $\Delta = 0.25$ ($z = 0.33$), where $\Psi = 3\sqrt{3}\ (64\Omega_0)^{-1/2} = 0.65\Omega_0^{-1/2}$, in the case $P = 0$. Similarly, when $P = \epsilon/3$, the maximum occurs at the value $\Delta = 1 - 2^{-1/2} = 0.29$ ($z = \sqrt{2} - 1 = 0.41$), where $\Psi = (4\Omega_0)^{-1/2}$.

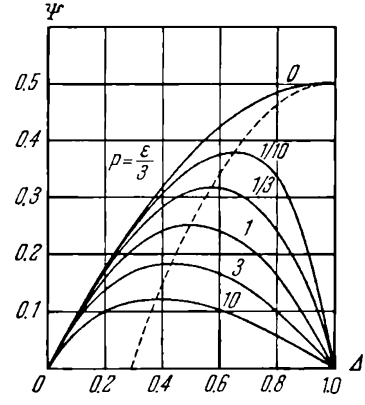


FIG. 9.—The dimensionless angular-diameter distance Ψ as a function of the redshift parameter Δ for various values of Ω and for the equation of state $P = \epsilon/3$.

For reference we present the following formulae:

$P = 0$;

$$\Psi_\Delta = \frac{2(1-\Delta)}{\Omega_0^2}\left[\left[2(\Omega_0 - 1)\Delta + (\Omega_0 - 2)\right.\right.$$
$$\left.\left. \times \{[1 + (\Omega_0 - 2)\Delta - (\Omega_0 - 1)\Delta^2]^{1/2} - 1\}\right]\right], \tag{3.3.9a}$$

$$\Psi_z = \frac{2}{\Omega_0^2(1+z)^2}\{\Omega_0 z + (\Omega_0 - 2)[(1 + \Omega_0 z)^{1/2} - 1]\} . \tag{3.3.9b}$$

$P = \epsilon/3$:

$$\Psi_\Delta = \frac{(1-\Delta)}{\Omega_0}\{[1 + 2(\Omega_0 - 1)\Delta - (\Omega_0 - 1)\Delta^2]^{1/2} - 1 + \Delta\} , \tag{3.3.10a}$$

$$\Psi_z = \frac{1}{\Omega_0(1+z)^2}\{[1 + \Omega_0 z(z+2)]^{1/2} - 1\} . \tag{3.3.10b}$$

The conclusions which follow from these relations will be discussed in the next section. The expansions for small Δ will be examined separately in § 3.5.

The observable luminosity of an object whose size, brightness, spectrum, and redshift are known is completely determined by the function $\Psi_\Delta(\Delta, \Omega_0)$. It is convenient to separate the question of the luminosity into a question about the brightness and one about the angular dimensions.

The brightness is measured by the amount of energy incident on an observer per unit receiver area and unit solid angle for rays incident perpendicular to the plane of the receiver: $I = d^2E/d\Xi dS$. By virtue of the isotropy of the Universe, an element of solid angle is expressed by spherical polar coordinates, in the usual way, namely, as $d\Xi = \sin\theta d\theta d\phi$. Moreover, in addition to the brightness integrated over the spectrum, one can introduce the spectral brightness distribution $d^3E = F(\omega)d\omega d\Xi dS$.

From optics it is known that in flat space, as long as no absorption or scattering of the light occurs, the brightness depends neither on distance nor on any refraction of the propagating light rays. For an expanding Universe, there is an analogous theorem: If, along the path of the light ray, there is neither scattering nor absorption, then the brightness depends on the redshift but not on the curvature of space or the density of matter. This was deduced in a very general form by Lindquist (1966) as a consequence of the Liouville theorem.

Now, radiation emitted by a blackbody with absolute temperature $T_{\rm em}$ is observed as radiation characteristic of a blackbody at the temperature $T_{\rm obs}$ $(1 - \Delta)T_{\rm em} = T_{\rm em}/(1 + z)$. Meanwhile, the Planck formula has the form

$$F(\omega, T) = \text{const. } \omega^3 f\left(\frac{\omega}{T}\right) d\omega = \text{const. } \frac{\omega^3 d\omega}{\exp(\hbar\omega/kT) - 1}, \qquad (3.3.11)$$

so the total flux of radiation in an element of solid angle is

$$I = \int_0^\infty F(\omega, T)d\omega = \text{const. } T^4. \qquad (3.3.12)$$

Consequently, I transforms with respect to redshift by the law

$$I = I_0(1 - \Delta)^4; \qquad (3.3.13a)$$

$$I = I_0(1 + z)^{-4}. \qquad (3.3.13b)$$

A spectrum of arbitrary form, characterized by the spectral distribution $F(\omega)$, will transform by the following more general law:

$$\begin{aligned} F(\omega_{\rm obs})d\omega_{\rm obs} &= (1 - \Delta)^4 F(\omega_{\rm em})d\omega_{\rm em} \\ &= (1 - \Delta)^3 F\left(\frac{\omega_{\rm obs}}{1 - \Delta}\right) d\omega_{\rm obs}; \end{aligned} \qquad (3.3.14)$$

where

$$\omega_{\rm obs} = (1 - \Delta)\omega_{\rm em}, \quad d\omega_{\rm obs} = (1 - \Delta)d\omega_{\rm em}, \qquad (3.3.14a)$$

$$\omega_{\rm obs} = \frac{\omega_{\rm em}}{1 + z}, \quad d\omega_{\rm obs} = \frac{d\omega_{\rm em}}{1 + z}, \qquad (3.3.14b)$$

so that the integrated (bolometric) brightness always decreases as $(1 - \Delta)^4 = (1 + z)^{-4}$, independent of the form of the source spectrum.

The differential law for the change in the spectral brightness at a given, fixed frequency depends on the form of the emission spectrum. The formula for an arbitrary emission spectrum was presented above. In the particular case of a power-law spectrum,

$$F(\omega_{\rm em})d\omega_{\rm em} = A\omega_{\rm em}{}^n d\omega_{\rm em} , \tag{3.3.15}$$

the observed spectrum will maintain a power-law form:

$$\begin{aligned} F(\omega_{\rm obs})d\omega_{\rm obs} &= A(1 - \Delta)^{3-n}\omega_{\rm obs}{}^n d\omega_{\rm obs} \\ &= A(1 + z)^{n-3}\omega_{\rm obs}{}^n d\omega_{\rm obs} . \end{aligned} \tag{3.3.16}$$

Having explained the law for the change in brightness, i.e., the flux of light per unit solid angle, let us turn to the question of the total amount of light falling on an observer from a given source.

The angular size of a distant star cannot be resolved. Therefore, its surface brightness is not measured but, rather, the flux of radiation from the entire star incident per unit area is the observed quantity.

If Euclidean geometry were applicable, the solid angle subtended by a source would be equal to the ratio of the projection of the area of the source (on the plane perpendicular to the line of sight) to the square of its distance away:

$$\Xi = S/R^2 \quad \text{(Euclidean)} . \tag{3.3.17}$$

In the expanding Universe we must replace R by $\bar{R}$ so that

$$\Xi = S \Big/ \left(\frac{c}{H_0}\Psi\right)^2 , \tag{3.3.18}$$

where the definition of Ψ is given above.

The total amount of light incident on a unit surface area at the observer is expressed in terms of the luminosity of the star $L_{\rm em}$:

$$\mathfrak{F}_{\rm obs} = \frac{L_{\rm em}}{4\pi(c/H_0)^2}\frac{(1 - \Delta)^4}{\Psi_\Delta{}^2} ; \tag{3.3.19a}$$

$$\mathfrak{F}_{\rm obs} = \frac{L_{\rm em}}{4\pi(c/H_0)^2}\frac{1}{\Psi_z{}^2(1 + z)^4} . \tag{3.3.19b}$$

One can define the "luminosity" distance D as the distance at which a motionless body in Euclidean space would produce an equal flux of energy at the observer. From the equations

$$\mathfrak{F}_{\rm obs} \equiv \frac{L_{\rm em}}{4\pi D^2} = \frac{L_{\rm em}(1 - \Delta)^4}{4\pi(c/H_0)^2\Psi_\Delta{}^2} , \tag{3.3.20a}$$

$$\mathfrak{F}_{\rm obs} = \frac{L_{\rm em}}{4\pi(c/H_0)^2\Psi_z{}^2(1 + z)^4} , \tag{3.3.20b}$$

and

$$D = \frac{c}{H_0} \frac{\Psi_\Delta(\Delta, \Omega_0)}{(1 - \Delta)^2}, \tag{3.3.21a}$$

$$D = \frac{c}{H_0} (1 + z)^2 \Psi_z(z, \Omega_0) ; \tag{3.3.21b}$$

for small Δ we have $\Psi_\Delta(\Delta, \Omega_0) \approx \Delta$, and therefore

$$D \approx \frac{c}{H_0} \Delta \approx \frac{c}{H_0} z \approx \bar{R} ,$$

i.e., naturally, all the different definitions of distance coincide in this limit. For Δ comparable with unity, the denominator $(1 - \Delta)^2$ plays a large role. With increasing Δ, the quantity D grows monotonically and does not have a maximum, in distinction to $\bar{R}$. In all variants of the model ($P = 0$, $P = c^2/3$; and for any Ω_0) $D \to \infty$ for $\Delta \to 1$. This means that the redshift substantially decreases the flux of energy from distant stars.

In reality what is important is the flux of energy in the part of the spectrum which can be registered by a photographic plate or some other type of device sensitive to radiation. This consideration leads one to the study of spectral corrections. The resulting theory of spectral corrections (the so-called K-correction), which take account of the receiver sensitivity, is based on formulae (3.3.14) and is well developed (McVittie 1959*a*, 1962*a*, *b*; Whitford 1971; Oke 1967; Sandage 1968); we will not present it here.

In expression (3.3.19) $\mathcal{F}_{\text{obs}}$ is the energy received per unit of observer proper time, whereas L_{em} is the energy produced per unit source proper time. Here, as above, the transformation of time is indeed accounted for; if a time τ elapses between two events (for example, two flares) at the star, then the observer will receive the light signals from these events in a larger time interval $\tau_{\text{obs}} = \tau/(1 - \Delta) = \tau(1 + z)$.

Finally it remains for us to discuss one other type of observation, namely that connected with the amount of matter contained in the layer between the redshifts Δ and $\Delta + d\Delta$ or z and $z + dz$. We shall again begin with the classical picture. In a Euclidean, static Universe, it is evident that

$$dM = \rho_0 4\pi R^2 dR , \tag{3.3.22}$$

or, if we speak about the number of particles (for example, nucleons),

$$dN = n_0 4\pi R^2 dR . \tag{3.3.23}$$

In place of R we substitute the expression in terms of Δ. For small Δ, $R \approx (c/H_0)\Delta$ and we obtain

$$dN \approx n_0 4\pi (c^3/H_0^3)\Delta^2 d\Delta , \tag{3.3.24}$$

and an analogous form for dM.

We will now introduce the function $\zeta_\Delta(\Delta, \Omega_0)$ which "corrects" the classical formula; for an expanding, curved Universe we have

$$dN = n_0 \frac{4\pi c^3}{H_0^3} \zeta_\Delta(\Delta, \Omega_0)\Delta^2 d\Delta \,. \qquad (3.3.25)$$

This function is drawn for the two cases $P = 0$ (Fig. 10) and $P = \epsilon/3$ (Fig. 11). The function ζ_Δ varies within wide limits; therefore, for each case, separate graphs (Figs. 12 and 13) are drawn using a large scale so that it is convenient to read off the value of ζ_Δ for small $\Delta \ll 1$. According to the meaning of the correction it is clear that as $\Delta \to 0$ we have $\zeta_\Delta \to 1$ for any Ω_0. For small Ω_0 the curves for ζ_Δ achieve values which greatly exceed unity.

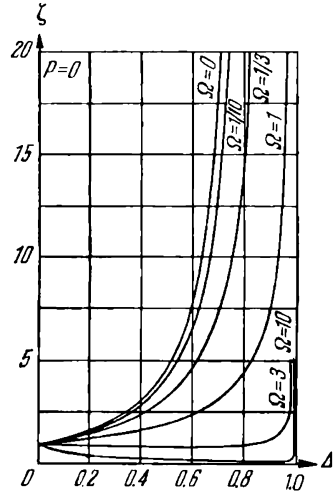


FIG. 10.—The dimensionless correction factor $\zeta(\Delta, \Omega)$ in the formula for the amount of material in a layer with a given redshift. The correction factor is given as a function of the redshift parameter Δ, for various values of Ω and for the equation of state $P = 0$.

If the variable z is used instead of Δ, we must give a new definition for the dimensionless correcting function; since

$$dN = n_0 \frac{4\pi c^3}{H_0^3} \zeta_z(z, \Omega_0) z^2 dz \,,$$

the new function ζ_z is connected with ζ_Δ by the relation

$$\zeta_z z^2 dz = \zeta_\Delta \Delta^2 d\Delta \,; \qquad \zeta_z = (1 - \Delta)^4 \zeta_\Delta = (1 + z)^{-4} \zeta_\Delta \,.$$

Therefore, in order to use the graphs of Figures 10–13 with z as a variable, one must transform the scale of the abscissa as well as that of the ordinate;

one must divide the function ζ_Δ by $(1 + z)^4$ in order to obtain $\zeta_z(z, \Omega_0)$. The qualitative behavior of ζ_Δ differs from that of ζ_z: the latter always has a negative derivative ($\partial\zeta_z/\partial z < 0$), so that ζ_z always decreases. For small z for both cases $P = 0$ and $P = \epsilon/3$, we have

$$\zeta_z \approx 1 - 2z(1 + \Omega_0) .$$

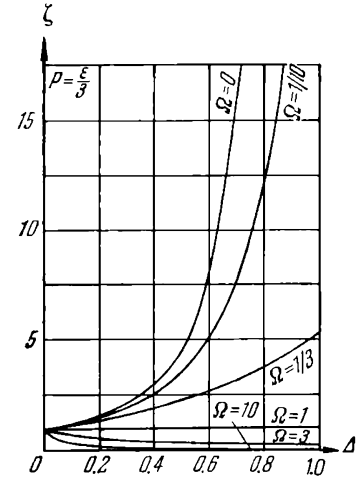


FIG. 11.—The dimensionless correction factor $\zeta(\Delta, \Omega)$ for the equation of state $P = \epsilon/3$

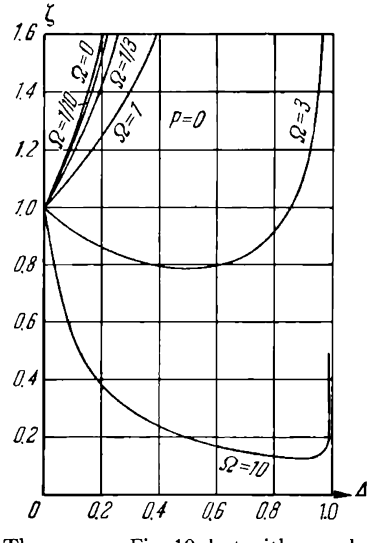


FIG. 12.—The same as Fig. 10, but with an enlarged scale

The curves are for $\Omega_0 = 0, 1/10, \frac{1}{3}, 1, 3, 10$. For $\Omega_0 = 0$ (the Milne model), there are simple formulae:

$$\zeta_\Delta = \frac{(2 - \Delta)^2}{4(1 - \Delta)^3}, \tag{3.3.26a}$$

$$\zeta_z = \frac{(2 + z)^2}{4(1 + z)^3}. \tag{3.3.26b}$$

For $P = 0$, $\Delta \to 1$: $\zeta_\Delta \to \infty$ for any Ω_0, as is apparent from the diagrams. The strong dependence of ζ on Ω_0 may possibly be useful for an observational determination of Ω_0 (in this connection see § 3.9 and § 3.10). The

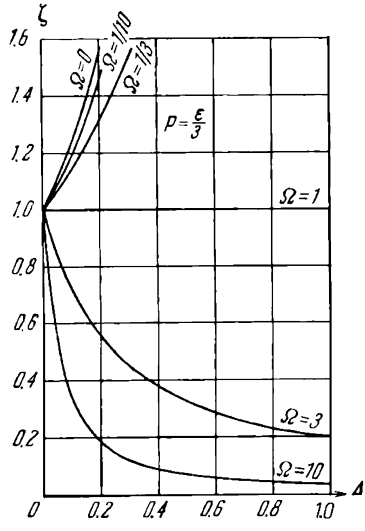


FIG. 13.—The same as Fig. 11, but with an enlarged scale

distribution of objects with respect to observable quantities may also be expressed by means of the function ζ; we shall be concerned with this in § 3.5.

For practical applications of formula (3.3.25), the assumption is often made that the number of observed objects of a definite type—for example, the number of galaxies in a given layer—is proportional to the total number of nucleons in this layer. In the case $P = 0$ one can say more simply that the number of galaxies is proportional to the total mass in the given layer.

The case $P = \epsilon/3 = \rho c^2/3$ corresponds to the assumption that photons and neutrinos dominate the total density ρ, while ordinary matter, of which the stars are made, contributes a small fraction of the total density. This fraction necessarily varies, since in the course of the expansion a change in the energy of the photons and neutrinos occurs. From the law for

the adiabatic expansion of matter whose equation of state is $P = \rho c^2/3$, the number density of nucleons will satisfy

$$n = \text{const.}\ \rho^{3/4}\,. \tag{3.3.27}$$

In this case, one naturally assumes that the density of galaxies and other such objects is proportional to n, and not to ρ. The correction function ζ is indeed given for dN and n, not for dM and ρ.

For reference purposes, we present the general formulae for ζ, which, unfortunately, are very cumbersome and inconvenient; we also specialize to some particular cases.

For $P = 0$:

$$\zeta(\Delta; \Omega_0) = 1 - \frac{2(\Omega_0 - 1)\Delta + (\Omega_0 - 2)\{[1 + (\Omega_0 - 2)\Delta - (\Omega_0 - 1)\Delta^2]^{1/2} - 1\}}{\Omega_0^2\,\Delta^2[1 + (\Omega_0 - 2)\Delta - (\Omega_0 - 1)\Delta^2]^{1/2}}, \tag{3.3.28a}$$

$$\zeta_z(z; \Omega_0) = \frac{4}{\Omega_0^4}\frac{[\Omega_0 z + (\Omega_0 - 2)(1 + \Omega_0 z)^{1/2} - 1)]^2}{z^2(1 + z)^3(1 + \Omega_0 z)^{1/2}}. \tag{3.3.28b}$$

The asymptotic limits of these formulae are:

$$\Delta \ll 1: \quad \zeta_\Delta(\Delta; \Omega_0) \approx 1 - (\Omega_0 - 2)\Delta\,, \tag{3.3.29a}$$

$$z \ll 1: \quad \zeta_z(z; \Omega_0) \approx 1 - (\Omega_0 + 2)z\,; \tag{3.3.29b}$$

$$1 - \Delta \ll 1: \quad \zeta_\Delta(\Delta; \Omega_0) \approx \frac{4}{\Omega_0^{5/2}}\frac{1}{(1 - \Delta)^{1/2}}\,, \tag{3.3.30a}$$

$$z \gg 1: \quad \zeta_z(z; \Omega_0) \approx \frac{4}{\Omega_0^{5/2}}\frac{1}{z^{7/2}}\,. \tag{3.3.30b}$$

For the particular value $\Omega_0 = 0$ we have (3.3.26); for $\Omega_0 = 1$ we obtain:

$$\zeta_\Delta(\Delta; 1) = \frac{4}{\Delta^2}\frac{[1 - (1 - \Delta)^{1/2}]^2}{(1 - \Delta)^{1/2}}\,; \tag{3.3.31a}$$

$$\zeta_z(z; 1) = \frac{4}{z^2}\frac{[(1 + z)^{1/2} - 1]^2}{(1 + z)^{5/2}}\,. \tag{3.3.31b}$$

For $P = \epsilon/3$ we have the results:

$$\zeta_\Delta(\Delta; \Omega_0) = \frac{1}{\Omega_0^2}\frac{1}{\Delta^2}\frac{\{[(1 - \Omega_0)(1 - \Delta)^2 + \Omega_0]^{1/2} - (1 - \Delta)\}^2}{[(1 - \Omega_0)(1 - \Delta)^2 + \Omega_0]^{1/2}}\,; \tag{3.3.32a}$$

$$\zeta_z(z; \Omega_0) = \frac{1}{\Omega_0^2}\frac{1}{z^2}\frac{\{[(1 - \Omega_0) + \Omega_0(1 + z)^2]^{1/2} - 1\}^2}{(1 + z)^3[(1 - \Omega_0) + \Omega_0(1 + z)^2]^{1/2}}. \tag{3.3.32b}$$

The asymptotic limits of these formulae are:

$$\Delta \ll 1: \quad \zeta_\Delta(\Delta; \Omega_0) \approx 1 - 2(\Omega_0 - 1)\Delta\,, \tag{3.3.33a}$$

$$z \ll 1 : \quad \zeta_z(z; \Omega_0) \approx 1 - 2(\Omega_0 + 1)z ; \tag{3.3.33b}$$

$$1 - \Delta \ll 1 : \quad \zeta_\Delta(\Delta; \Omega_0) \approx \frac{1}{\Omega_0^{3/2}} ; \tag{3.3.34a}$$

$$z \gg 1 : \quad \zeta_z(z; \Omega_0) \approx \frac{1}{\Omega_0^{3/2}} \frac{1}{z^4} . \tag{3.3.34b}$$

Equation (3.3.26) applies for $\Omega_0 = 0$ (independent of the equation of state). For $\Omega_0 = 1$ and $P = \epsilon/3$ we have:

$$\zeta_\Delta(\Delta; 1) = 1 , \tag{3.3.35a}$$

$$\zeta_z(z; 1) = \frac{1}{(1 + z)^4} . \tag{3.3.35b}$$

The functions ζ are simply related to the functions g introduced in § 3.2:

$$g(\Omega_0) = 3 \int_0^1 \zeta_\Delta(\Delta; \Omega_0)\Delta^2 d\Delta = 3 \int_0^\infty \zeta_z(z; \Omega_0) z^2 dz . \tag{3.3.36}$$

Although $\zeta_\Delta(\Delta; \Omega_0) \to \infty$ as $\Delta \to 1$ when $P = 0$, the integral is convergent except when $\Omega_0 = 0$.

Finally, we call attention to a convenient formula connecting dt with dz (valid for $P = 0$):

$$H_0 dt = \frac{dz}{(1 + z)^2(1 + \Omega_0 z)^{1/2}} = \left[\frac{1 - \Delta}{1 + (\Omega_0 - 1)\Delta}\right]^{1/2} d\Delta . \tag{3.3.37}$$

3.4 Working Formulae as Functions of the Redshift z

In this section we will obtain formulae which are often used directly to interpret the observational data.

It is believed that during the recent epochs characterized by the existence of discrete objects such as galaxies and quasars, the pressure, P, is negligible in comparison with the density ρc^2. So, employing this assumption, let us rewrite formulae (2.1.8) and (2.1.10) using the notation $H(t) = \dot{a}(t)/a(t)$:

$$\frac{dH(t)}{dt} + H^2(t) = -4\pi G\rho(t)/3 ; \tag{3.4.1a}$$

$$\frac{d\rho(t)}{dt} = -3\rho(t)H(t) . \tag{3.4.1b}$$

Dividing, we obtain

$$\frac{dH(t)}{d\rho(t)} = \frac{4\pi G\rho(t) + 3H^2(t)}{9\rho(t)H(t)} . \tag{3.4.2}$$

It is convenient to introduce $\mathfrak{h} = \mathfrak{h}(t) \equiv H(t)/H_0$ ($\mathfrak{h}$ is not to be confused with the previously introduced h_0 where $H_0 = 100h_0$ km s^{-1} Mpc^{-1}) and recall that

$$(1 + z)^3 \quad \frac{\rho(t)}{\rho_0} = \frac{\rho(t)}{\rho_c} \frac{\rho_c}{\rho_0} = \frac{8\pi G\rho(t)}{3H_0^2\Omega_0} .$$

Remember that H_0 is the present value of the Hubble parameter, ρ_0 is the present value of the matter density, and Ω_0 is the present ratio of the density to the critical density, ρ_c, the latter calculated using the present value of $H(t_0) \equiv H_0$.

The quantity z is the redshift of the light emitted at the time t. This quantity is a convenient parameter for the solution of equations and, indeed, is the directly observed quantity used in research on distant objects.

In terms of these new variables, we find after some straightforward algebra that

$$\frac{d\mathfrak{h}}{dz} = \frac{\mathfrak{h}}{1+z} + \frac{\Omega_0(1+z)^2}{2\mathfrak{h}}. \tag{3.4.3}$$

The initial conditions are in essence an identity: $z = 0$ and $\mathfrak{h} = 1$ at the present. The equation is easy to integrate; it is homogeneous and the variables separate via the substitution

$$\mathfrak{C} \equiv \mathfrak{h}^2(1+z)^{-3}, \tag{3.4.4a}$$

to give

$$\frac{d\mathfrak{C}}{dz} = \frac{(\Omega_0 - \mathfrak{C})}{(1+z)} \tag{3.4.4b}$$

The solution is

$$\mathfrak{h} = (1+z)(1+\Omega_0 z)^{1/2}. \tag{3.4.5}$$

From the equation for $\rho(t)$ we obtain

$$\frac{d[(1+z)^3]}{dt} = -3H_0\mathfrak{h}(1+z)^3; \qquad H_0 dt = -\frac{dz}{\mathfrak{h}(1+z)}. \tag{3.4.6}$$

Note that this equation also follows directly from the connections between the scale, the redshift, and the Hubble parameter:

$$a(t) = \frac{a_0}{1+z}; \qquad \frac{da(t)}{dt} = -H(t)a(t) = -\mathfrak{h}H_0 a(t).$$

Next substituting for $\mathfrak{h}$, we find

$$H_0 dt = \frac{-dz}{(1+z)^2(1+\Omega_0 z)^{1/2}}. \tag{3.4.7}$$

The redshift (z) is often a more convenient parameter than the time (t). An integral over time can be transformed into an integral over redshift by using (3.4.7) and expressing all quantities in terms of z.

With the help of equation (3.4.7), it is not difficult to find t as a function of z. The time t, from the singularity ($z = \infty$, $t = 0$), is

$$H_0 t = \int_z^\infty \frac{dz'}{(1+z')^2(1+\Omega_0 z')^{1/2}}. \tag{3.4.8}$$

This integral can be done exactly, giving the analytic forms

$$t = \frac{H_0^{-1}}{(1-\Omega_0)}\left\{\frac{(1+\Omega_0 z)^{1/2}}{(1+z)} + \frac{\Omega_0}{2(1-\Omega_0)^{1/2}} \times \ln\left[\frac{(1+\Omega_0 z)^{1/2}-(1-\Omega_0)^{1/2}}{(1+\Omega_0 z)^{1/2}+(1-\Omega_0)^{1/2}}\right]\right\}, \quad \Omega_0 < 1;$$

$$t = \frac{H_0^{-1}}{(\Omega_0-1)}\left\{\frac{\pi}{2}\frac{\Omega_0}{(\Omega_0-1)^{1/2}} - \frac{\Omega_0}{(\Omega_0-1)^{1/2}}\tan^{-1}\left[\left(\frac{1+\Omega_0 z}{\Omega_0-1}\right)^{1/2}\right] - \frac{(1+\Omega_0 z)^{1/2}}{(1+z)}\right\}, \quad \Omega_0 > 1; \tag{3.4.9}$$

$$t = \frac{2H_0^{-1}}{3(1+z)^{3/2}}, \quad \Omega_0 = 1.$$

For the general case when $\Omega_0 \neq 1$, $\Omega_0 \neq 0$, the integral cannot be inverted, and one cannot give an analytic expression for z in terms of t.

The value of the integral for $z = 0$ (the present age of the Universe) was given in formulae (1.3.1) and (1.3A.2) above; also, Figure 3 presents the curve giving the dependence of t_0 (i.e., $z = 0$) on Ω_0.

It is easy to find the asymptotic forms when $z \gg 1$, $\Omega_0 z \gg 1$ from the integral (3.4.8), but not from expression (3.4.9). We obtain

$$t = \frac{2}{3H_0\Omega_0^{1/2}z^{3/2}}. \tag{3.4.9a}$$

From this form, all the well-known results for small t and large z are apparent. From equations (3.4.1), (3.4.5), and (3.4.9a), we have in this limit

$$H(t) \approx \frac{2}{3t}, \quad \rho(t) \approx \frac{1}{6\pi G t^2}. \tag{3.4.10}$$

It is also easy to express the instantaneous value of $\Omega(t)$ in terms of z and the present value Ω_0:

$$\Omega(t) = \Omega_0\frac{\rho(t)}{\rho_0}\frac{H_0^2}{H^2(t)} = \frac{\Omega_0(1+z)^3}{\mathfrak{h}^2} = \Omega_0\frac{(1+z)^3}{(1+z)^2(1+\Omega_0 z)} = \frac{\Omega_0(1+z)}{(1+\Omega_0 z)} = 1 - \frac{1-\Omega_0}{1+\Omega_0 z}. \tag{3.4.11}$$

This confirms the result that all local properties approximate the properties of the critical density model as the singularity is approached (i.e., $\Omega(z) \to 1$ for $\Omega_0 z \gg 1$).

The parameter z is also convenient for obtaining formulae relating to the observational quantities—the angular diameter of distant bodies and the luminosity distance (these formulae were given without derivation in § 3.3).

First, we may express the present radius of the Universe in terms of H_0 and Ω_0. We develop the appropriate expression for the open model, $\Omega_0 < 1$

(the calculations are easily repeated for the closed model $\Omega_0 > 1$). For $\Omega_0 < 1$ I recall that

$$a_0 = \frac{c}{H_0(1 - \Omega_0)^{1/2}}, \tag{3.4.12}$$

at the present, while at any arbitrary time

$$a(z) = \frac{a_0}{1 + z} = \frac{c}{H_0(1 + z)(1 - \Omega_0)^{1/2}}. \tag{3.4.13}$$

Let us express the dimensionless, comoving coordinate r of the observed galaxy through z; i.e., let us find r for a galaxy having the redshift z. From the condition $ds^2 = c^2dt^2 - a^2dr^2$, we find

$$dr = \frac{-cdt}{a} = \frac{(1 - \Omega_0)^{1/2}dz}{(1 + z)(1 + \Omega_0 z)^{1/2}}. \tag{3.4.14}$$

The corresponding integral is expressible via elementary functions:

$$(1 - \Omega_0)^{1/2} \int_0^z \frac{dz'}{(1 + z')(1 + \Omega_0 z')^{1/2}} = \sinh^{-1}\left[\left[\frac{2(1 - \Omega_0)^{1/2}}{\Omega_0^2(1 + z)}\{\Omega_0 z + (2 - \Omega_0)[1 - (1 + \Omega_0 z)^{1/2}]\}\right]\right]. \tag{3.4.15}$$

The observed bolometric (total) flux from a source is determined by the amount of energy incident on a unit area per unit time. For a source with comoving coordinate r and absolute luminosity L_{em}, this is

$$\mathcal{F}_{obs} = \frac{L_{em}}{[4\pi a_0^2 \sinh^2 r][(1 + z)^2]}. \tag{3.4.16}$$

The first factor in the denominator is the area of the sphere, whose comoving coordinate is r, at the moment the light is received (t_0). The second factor accounts for the decrease in intensity due to the redshift. This formula may be cast in the form

$$\mathcal{F}_{obs} = \frac{L_{em}}{4\pi\bar{R}^2(1 + z)^4}. \tag{3.4.17}$$

Here, the quantity $4\pi\bar{R}^2$ in the denominator enters because $\mathcal{F}_{obs}$ is inversely proportional to the solid angle subtended by the source; the factor $(1 + z)^4$ gives the change in brightness (which is proportional to T^4, the energy per unit solid angle; see § 3.3).

Another interpretation is possible. The quantity $4\pi a_0^2 \sinh^2 r$ is the area of the sphere over which the radiation is spread at the present time [its radius equals $\bar{R}(1 + z)$]. In addition we must divide by $(1 + z)^2$ to take account of the decrease in the energy of the photons and the decrease in the rate at which they are received.

Astronomers use the apparent magnitude m (the logarithm of the observed flux) for reporting measurements. By definition,

$$m = -2.5 \log_{10} \mathcal{F}_{obs} + \text{const}. \tag{3.4.18}$$

The absolute magnitude M of an object is the apparent magnitude this object would have if it were located 10 parsecs ($= 3 \times 10^{19}$ cm) away.

Let us now cast (3.4.16) in a form using apparent magnitudes (3.4.18); we substitute expansions (3.4.12) and (3.4.15) and the deceleration parameter $q_0 = \frac{1}{2}\Omega_0$ (see eq. [1.2.6]) to obtain

$$m_{\text{bol}} = 5 \log_{10} \frac{1}{q_0^2} \{q_0 z + (q_0 - 1)[(1 + 2q_0 z)^{1/2} - 1]\} + C_1 . \quad (3.4.19)$$

The quantity C_1 depends only on the absolute magnitude M of the source (or, equivalently, L_{em}) and on H_0. From the definition of M we obtain

$$C_1 = M_{\text{bol, at emission}} - 45.06 - 5 \log_{10} H_0 . \quad (3.4.20)$$

Here, H_0 is expressed in s^{-1}. Formula (3.4.19) is also valid for the closed model, i.e., when $q_0 > \frac{1}{2}$. This formula was obtained by Mattig (1958). It is the basis in cosmology for the analysis of the relationship between apparent magnitudes and redshifts.

In the following section, we will obtain another formula, suitable for the analysis of objects with small redshift ($z \lesssim 0.3$).

3.5 The First Approximation and Euclidean Space; the Deceleration Parameter

In the above, we have systematically compared the general results obtained for different cosmological models with the simplest classical picture of the Universe. In the latter, there is a unique meaning for the distance R. The amount of light an observer receives from a motionless source decreases as R^{-2}; the volume of space grows as R^3; and the total density of matter in no way influences the structure of space.

Exact formulae applying to the expanding Universe depend both on the density and the equation of state of the matter; in particular, formulae were given above for $P = 0$ and $P = \epsilon/3$.

It is clear that through experimental research, very precise data may be obtained about nearby regions. For the interpretation of these data, it is natural to expand the theoretical formulae in powers of Δ or z. In such an expansion, the first term must necessarily coincide with the classical result while the subsequent terms represent the corrections to the classical picture.

But the corrections, which depend on the curvature of space, are proportional to the square of the distance. These corrections must also be proportional to the density, and so in order to obtain a dimensionless quantity, the combination $G\rho R^2/c^2$ is used. The distance is proportional to the redshift; consequently, the corrections in question are of order Δ^2, which, when applied to quantities which are proportional to Δ, gives terms of order Δ^3. From this it follows that the formulae which take account of

terms $\sim \Delta^2$ (or $\sim \varepsilon^2$) can be systematically obtained without considering the structure of space. We shall consider the case $P = 0$.

Let us examine motion in Euclidean space and laboratory time using the Newtonian theory of gravity. We choose the system of coordinates so that the observer is located at the origin and is at rest. We take as the zero of time the present moment (the moment of observation) and not, as before, the moment of the singularity $\rho = \infty$. Consider the case $P = 0$. For $t = 0$, the matter density is a fixed constant over all space and the matter moves according to the Hubble law $u = H_0 R$. The difference between laboratory time and proper time is of order Δ^2 and hence the statements regarding Hubble's law or constant density are only true to the order needed for subsequent estimates.

Now write the hydrodynamical equation of motion and the continuity equation:

$$\frac{\partial \rho}{\partial t} \qquad \operatorname{div}(\rho u) = -3H\rho\,, \qquad \frac{\partial u}{\partial t} + u\frac{\partial u}{\partial R} = -\frac{G4\pi\rho R^3}{3R^2}\,. \tag{3.5.1}$$

The first terms in a series expansion of the solution for small t are

$$\rho \approx \rho_0 - 3H_0 t\rho_0\,, \qquad u \approx H_0 R - H_0^2 R t - \frac{4\pi}{3} G\rho_0 R t\,. \tag{3.5.2}$$

Recalling the definitions $\rho_c \equiv 3H_0^2/(8\pi G)$ and $\Omega_0 = \rho_0/\rho_c$, we can rewrite the second formula above as

$$u \approx H_0 R - H_0^2(1 + \Omega_0/2)Rt\,. \tag{3.5.3}$$

Now we shall turn to the propagation of light and the redshift. A photon emitted at time t ($t < 0$) at a distance R will reach the origin at $t = 0$ if

$$R = -ct\,.^1 \tag{3.5.4}$$

If the photon is emitted by a particle moving with velocity u, the frequency in the laboratory system will be ω', where

$$\frac{\omega'}{\omega_{\text{em}}} = 1 - \Delta' = \left(\frac{1 - u/c}{1 + u/c}\right)^{1/2}. \tag{3.5.5}$$

We wish to include corrections of order Δ'; we therefore expand the above expression, keeping only the appropriate terms:

$$\Delta' \approx \frac{u}{c} - \frac{1}{2}\left(\frac{u}{c}\right)^2. \tag{3.5.6}$$

1. If τ is the proper time, then $t \neq \tau_{\text{obs}} - \tau_{\text{em}}$ and the difference is of order Δ^2, which is the same order as the effects we are taking account of. This was pointed out in a discussion between the authors and the editor (Gary Steigman) to whom we are grateful. However, the following discussion is not thereby invalidated. In particular, the second order corrections in observable quantities do not depend on space curvature.

We still must account for the frequency change due to the motion through the gravitational field. This shift yields:

$$\omega_{\text{obs}} - \omega' = \left[\frac{\Phi(r) - \Phi(0)}{c^2}\right]\omega' , \tag{3.5.7}$$

where Φ is the Newtonian gravitational potential. From Poisson's equation:

$$\nabla^2\Phi = 4\pi G\rho , \tag{3.5.8}$$

it follows for spherical symmetry that:

$$\Phi = \text{const.} + \frac{2\pi}{3} G\rho R^2 , \tag{3.5.9}$$

To the present level of approximation: $\rho \approx \rho_0$ and $\omega_{\text{obs}} \approx \omega'$, so that (3.5.7) may be rewritten as:

$$\omega_{\text{obs}} - \omega' \approx \frac{2\pi}{3c^2} G\rho_0 R^2 \omega_{\text{obs}} . \tag{3.5.7'}$$

Combining (3.5.5), (3.5.6), and (3.5.7'), we finally obtain

$$\Delta \approx \Delta' - \frac{2\pi}{3c^2} G\rho_0 R^2 \approx \frac{u}{c} - \frac{1}{2}\left(\frac{u}{c}\right)^2 - \frac{2\pi}{3c^2} G\rho_0 R^2 . \tag{3.5.10}$$

If we employ the expression (3.5.3) for the velocity, use (3.5.4) to express the time in terms of the distance, and write $\Omega_0 = 8\pi G\rho_0/3H_0^2$, we obtain to order R^2:

$$\Delta \approx \left(\frac{R}{c/H_0}\right) + \left(\frac{R}{c/H_0}\right)^2 \left(\frac{1}{2} + \frac{\Omega_0}{4}\right) . \tag{3.5.11}$$

We may solve this equation for R to order Δ^2 (or z^2):

$$R \approx \frac{c}{H_0}\left[\Delta - \left(\frac{1}{2} + \frac{\Omega_0}{4}\right)\Delta^2\right] , \tag{3.5.12a}$$

$$R \approx \frac{c}{H_0}\left[z - \left(\frac{3}{2} + \frac{\Omega_0}{4}\right) z^2\right] . \tag{3.5.12b}$$

This expression for R coincides with the first two terms in the series expansion of formula (3.3.9).

Correspondingly, for the time of emission of the light which will arrive at the observer at $t = 0$, we find

$$t \approx -\frac{1}{H_0}\left[\Delta + \left(\frac{1}{2} - \frac{\Omega_0}{4}\right)\Delta^2\right] , \tag{3.5.13a}$$

$$t \approx -\frac{1}{H_0}\left[z - \left(\frac{1}{2} + \frac{\Omega_0}{4}\right) z^2\right] . \tag{3.5.13b}$$

Remember that R is the source distance at the instant of emission t; the light, with redshift Δ, is received at $t = 0$, at the origin. In deriving (3.5.13) it is necessary to use in place of (3.5.4) an expression which is valid to order Δ^2: $-ct \approx R(1 + H_0R/c)$.

In our Euclidean approximation the distance has only one meaning. In that framework we find an obvious form for the angular size of a source of absolute linear size l. Thus, the "angular diameter" distance is R, so that

$$\theta = \frac{l}{R} \approx \frac{lH_0}{c}\frac{1}{\Delta}\left[1 + \left(\frac{1}{2} + \frac{\Omega_0}{4}\right)\Delta\right], \tag{3.5.14a}$$

$$\theta = \frac{l}{R} \approx \frac{lH_0}{c}\frac{1}{z}\left[1 + \left(\frac{3}{2} + \frac{\Omega_0}{4}\right)z\right]. \tag{3.5.14b}$$

The correction to the elementary formulae $\theta = lH_0/(c\Delta)$, $R = c\Delta/H_0$ is connected only with the corrections to the linear relations between the distance (at the moment of light emission) and the redshift.

From (3.1.17) we may write for the observed flux:

$$\mathcal{F}_{\text{obs}} = \frac{L_{\text{em}}}{4\pi D^2} = \frac{L_{\text{em}}(1-\Delta)^4}{4\pi R^2}.$$

Expanding and keeping only terms of first order in Δ, we find

$$\mathcal{F}_{\text{obs}} \approx \frac{L_{\text{em}}}{4\pi(c/H_0)^2}\left(\frac{1}{\Delta}\right)^2\left[1 - \left(3 - \frac{\Omega_0}{2}\right)\Delta\right], \tag{3.5.15}$$

$$D = \frac{R}{(1-\Delta)^2} \approx \frac{c}{H_0}\,\Delta\left[1 + \left(\frac{3}{2} - \frac{\Omega_0}{4}\right)\Delta\right], \tag{3.5.16a}$$

$$D = R(1+z)^2 \approx \frac{cz}{H_0}\left[1 + \left(\frac{1}{2} - \frac{\Omega_0}{4}\right)z\right]. \tag{3.5.16b}$$

Recall that D is the "luminosity" distance (see § 3.3).

The above formula for D coincides with the first two terms of the expansion of the Mattig formula (3.4.19). Although expansion of (3.4.19) is easy, our purpose has been to call attention to the physical result that the curvature of space does not influence the first two terms. Note, however, that the gravitational force is included due to the term $\sim\Omega_0 \sim G\rho_0$; this is what we mean in asserting that Newtonian theory is valid in this approximation.

We will now find an expression for the mass M contained within the sphere of radius R at the moment of emission. Substituting t from equation (3.5.13) into equation (3.5.2), we obtain

$$\begin{aligned} \rho &\approx \rho_0(1 + 3\Delta), \\ M = \frac{4\pi}{3}\rho R^3 &\approx \frac{4\pi}{3}\rho_0\left(\frac{c}{H_0}\Delta\right)^3\left[1 + \left(\frac{3}{2} - \frac{3}{4}\Omega_0\right)\Delta\right]. \end{aligned} \tag{3.5.17}$$

Now it is easy to compute the derivative $dM/d\Delta$:

$$\frac{dM}{d\Delta} \approx 4\pi\rho_0\left[\frac{c}{H_0}\right]^3[1 + (2 - \Omega_0)\Delta]\Delta^2. \tag{3.5.18}$$

The quantity in the second set of brackets is the function $\zeta_\Delta(\Delta, \Omega_0)$ of the previous section (more precisely, the first two terms of its Taylor series expansion),

$$\zeta_\Delta \approx 1 + (2 - \Omega_0)\Delta \,. \tag{3.5.19}$$

These formulae give all the information that is required for an interpretation of the observations when Δ is not too large ($\Delta \lesssim 0.3$). The corresponding formulae as a function of z are:

$$M \approx \frac{4\pi}{3}\rho_0\left(\frac{c}{H_0}z\right)^3\left[1 - \left(\frac{3}{2} + \frac{3\Omega_0}{4}\right)z\right], \qquad \zeta_z \approx 1 - (2 + \Omega_0)z \,. \tag{3.5.19a}$$

Previously in this section, only the case of negligible pressure, $P = 0$, has been considered. In the case of arbitrary pressure, both the gravitational field and the acceleration due to the force of gravity depend on the combination $\rho + 3P/c^2$, whereas the expression for the critical density does not change. Note that the formulae which result from the replacement of Ω_0 by $\Omega'_0 = 2\Omega_0$ in equations (3.5.12), (3.5.15), (3.5.16), and (3.5.19) coincide with the first terms in the series expansions of equations (3.3.10), (3.3.20), (3.3.21), and (3.3.33). Therefore for $P \neq 0$ it is necessary to substitute for Ω_0 the quantity

$$\Omega'_0 \equiv \Omega_0\left(1 + \frac{3P}{\rho c^2}\right)_0 \,. \tag{3.5.20}$$

In particular, $\Omega'_0 = 2\Omega_0$ for $P_0 = \rho_0 c^2/3$.

Correspondingly, analysis of the observations (for the range $\Delta \lesssim 0.3$) primarily reveals Ω'_0; in essence, we measure the deceleration caused by gravity. If $1 < \Omega'_0 < 2$, then, for a Universe filled with ordinary matter, $P = 0$, $\Omega_0 = \Omega'_0 > 1$, and the Universe is closed. But for a Universe whose dominant components are photons and neutrinos, $\Omega_0 = \Omega'_0/2 < 1$ and the Universe is open. In many books and reviews the Universe is characterized by two parameters:

$$\left(\frac{\dot{l}_{ab}}{l_{ab}}\right)_0 = h_1 \,, \quad \left(\frac{\ddot{l}_{ab}}{l_{ab}}\right)_0 = h_2 \,. \tag{3.5.21}$$

Since the distance between any pair of objects is proportional to the scale $a(t)$, the formulae (3.5.21) are equivalent to the formulae

$$\left(\frac{\dot{a}}{a}\right)_0 = h_1 \,, \quad \left(\frac{\ddot{a}}{a}\right)_0 = h_2 \,. \tag{3.5.22}$$

For the critical density model ($\Omega_0 = 1$) we introduce $b(t)$ (see § 3.2); although $b(t)$ is determined up to an arbitrary constant factor, h_1 and h_2 have completely determined values even in the critical density case.

The quantity h_1 is none other than Hubble's constant: $h_1 = H_0$. The quantity h_2 depends on the density and satisfies

$$h_2 = -\frac{4\pi}{3} G\rho_0 \left(1 + \frac{3P_0}{\rho_0 c^2}\right). \tag{3.5.23}$$

In place of h_2, one sometimes introduces the dimensionless deceleration parameter, which can be expressed as

$$q_0 = -\frac{h_2}{h_1^2} = \tfrac{1}{2}\Omega_0 \left(1 + \frac{3P}{\rho c^2}\right)_0 = \frac{\Omega'_0}{2}. \tag{3.5.24}$$

Thus, when $P = 0$, the critical value $q_0 = \frac{1}{2}$ separates the closed models (for $q_0 > \frac{1}{2}$) from the open models ($q_0 < \frac{1}{2}$). When $P = \epsilon/3$, the critical value is $q_0 = 1$.

The conclusion to be drawn from this section is that the formulae linear in q_0 and containing corrections of order Δ (but not of order Δ^2) do not contain the effects of the curvature of space and can be consistently deduced in Euclidean space with the help of Newton's theory of gravity (if account is taken both of the frequency change of photons caused by gravity and of the exact formula for the Doppler effect).

In concluding the section, let us rewrite formula (3.5.15) in the form common to astronomers; i.e., let us rewrite the formula in terms of magnitudes and the parameter q_0:

$$m_{\text{bol}} \approx 5 \log z + 1.086(1 - q_0)z + C_1 ;$$

the expression for C_1 is given by formula (3.4.20). The formula here is valid for $z \lesssim 0.3$. It was obtained by Heckmann (1942), Robertson (1955), and McVittie (1956).

3.6 The Distribution with Respect to Apparent Magnitude

In the classical case, when one does not take account of the curvature of space, or the effect of redshift on brightness, or the evolutionary effect, one obtains well-known, exact relations for the distributions with respect to apparent magnitude, of objects of a given type.

Consider sources whose luminosity is L. We will denote the number density of these sources by n. The flux received by an observer from a source whose distance is r is

$$\mathcal{F} = \frac{L}{4\pi r^2}, \tag{3.6.1}$$

so that

$$r = \left[\frac{L}{4\pi\mathcal{F}}\right]^{1/2}. \tag{3.6.2}$$

The number of sources within a sphere of radius r (each of which will produce a flux greater than $\mathcal{F}$) equals

$$N = \frac{4\pi}{3} r^3 n = \text{const.}\ nL^{3/2}\mathcal{F}^{-3/2} . \tag{3.6.3}$$

The corresponding differential law is

$$dN = \text{const.}\ \mathcal{F}^{-5/2} d\mathcal{F} . \tag{3.6.4}$$

Using the magnitude $m = \text{const.} - 2.5 \log \mathcal{F}$ (where $\mathcal{F}$ is measured in ergs $\text{cm}^{-2}\,\text{s}^{-1}$), we find

$$dN = \text{const.}\ 10^{(0.6m)} dm . \tag{3.6.5}$$

This law does not change even if sources of different types are present. For the subsequent development, it is useful to reproduce here the elementary proof of this result. Thus, assume that in each volume element there is a definite distribution of sources with respect to absolute magnitude; between the luminosities L and $L + dL$ the number of sources per unit volume is

$$dn = n_0 W(L) dL . \tag{3.6.6}$$

The distribution is normalized $[\int_0^\infty W(L)dL = 1]$ so that $W(L)dL$ is the fraction of sources having luminosity L; the total number density of all sources equals n_0. For each group having a luminosity between L and $L + dL$, we find the law

$$d^2N = \text{const.}\ W(L) L^{3/2}\mathcal{F}^{-5/2} dL d\mathcal{F} . \tag{3.6.7}$$

The symbol d^2 signifies that the right-hand side is a double differential: $dLd\mathcal{F}$. If one carries out the integration over dL, it follows that:

$$dN = A\mathcal{F}^{-5/2} d\mathcal{F} , \tag{3.6.8}$$

where

$$A = \text{const.}\ n_0 \int_0^\infty L^{3/2} W(L) dL . \tag{3.6.9}$$

Thus, the apparent-magnitude distribution law does not change; rather, instead of a factor of the number density of objects with one luminosity we now use the quantity A, which depends on the distribution of luminosities $W(L)$. Note that the expression which gives the total amount of light emitted per unit volume will contain a related (but different) factor: $n_0 \int_0^\infty L W(L) dL$.

In an expanding Universe, a more complicated law is obtained for sources of given absolute magnitude since the number of such sources in a given sphere varies in a complicated, non-power-law fashion. Furthermore, for large distances, the light propagation time is also very long. Therefore, one must also account for the evolutionary effect, i.e., the change with time in the distribution of sources with respect to absolute magnitude. The complete formula is cumbersome and not illuminating. We will only present

the approximate formula, which can be obtained, as described above, from an examination of the accelerated motion of objects in flat space. In accordance with the material discussed earlier, we will limit the corrections everywhere to those of first order in Δ.

The distribution of sources with respect to luminosity is characterized by the function

$$dn = n(t)W(L, t)dL . \tag{3.6.10}$$

Here t is reckoned from the present time, so that at the instant of emission

$$t = -R/c ;$$

n is the total number density of all sources referred to Euclidean (not the comoving!) space. Assuming that sources (galaxies) are conserved, we obtain

$$n \approx n_0(1 - 3H_0t) . \tag{3.6.11}$$

The variation of W with time determines the effect of evolution. Recall that the observed flux (apparent magnitude) is related to the intrinsic luminosity (absolute magnitude) by an expression which includes the effect of the redshift:

$$\mathcal{F} = \frac{L(1 - \Delta)^4}{4\pi R^2} , \tag{3.6.12a}$$

$$\mathcal{F} = \frac{L}{4\pi R^2(1 + z)^4} . \tag{3.6.12b}$$

The fourth power of $(1 - \Delta)$ (or $1 + z$) appears because we have been considering bolometric quantities.

In general, however, instead of the bolometric intensity, the observed quantity is the flux in a frequency bandwidth $\Delta\nu_0$ at the observed frequency ν_0. The flux density $S(\nu_0)$ is the intensity per unit bandwidth; it is related to the emitted power per unit bandwidth $P(\nu_e)$ by:

$$S(\nu_0) = \frac{P(\nu_e)}{(1 - \Delta)} \frac{(1 - \Delta)^4}{4\pi R^2} = (1 - \Delta)^3 \frac{P[\nu_0/(1 - \Delta)]}{4\pi R^2} , \tag{3.6.13a}$$

$$S(\nu_0) = \frac{(1 + z)P(\nu_e)}{4\pi R^2(1 + z)^4} = \frac{P[(1 + z)\nu_0]}{(1 + z)^3} \frac{1}{4\pi R^2} . \tag{3.6.13b}$$

If $P(\nu)$ is of the form $P \sim \nu^n$, then:

$$S(\nu_0) = (1 - \Delta)^{3-n} \frac{P(\nu_0)}{4\pi R^2} = (1 + z)^{n-3} \frac{P(\nu_0)}{4\pi R^2} . \tag{3.6.14}$$

If $P(\nu)$ is of the Planck form, we find for $h\nu_0/kT \gtrsim 3$:

$$S(\nu_0) \approx (1 - \Delta)^{h\nu_0/kT} \frac{P(\nu_0)}{4\pi R^2} = (1 + z)^{-h\nu_0/kT} \frac{P(\nu_0)}{4\pi R^2} . \tag{3.6.15}$$

For a power law spectrum ($P \sim \nu^n$) we have to order Δ:

$$S(\nu_0) \approx \frac{P(\nu_0)}{4\pi(c/H_0)^2} \times \frac{1}{\Delta^2}\left[1 + \left(n - 2 + \frac{\Omega_0}{2}\right)\Delta\right], \quad (3.6.16a)$$

$$S(\nu_0) \approx \frac{P(\nu_0)}{4\pi(c/H_0)^2} \times \frac{1}{z^2}\left[1 + \left(n + \frac{\Omega_0}{2}\right)z\right]. \quad (3.6.16b)$$

If a flux density $S(\nu_0)$ is observed for a source whose power (per unit frequency interval) is $P(\nu_0)$, then the redshift is:

$$\Delta \approx \left[\frac{P(\nu_0)/S(\nu_0)}{4\pi(c/H_0)^2}\right]^{1/2}\left\{1 + \left(\frac{n}{2} - 1 + \frac{\Omega_0}{4}\right)\left[\frac{P(\nu_0)/S(\nu_0)}{4\pi(c/H_0)^2}\right]^{1/2}\right\}, \quad (3.6.17a)$$

$$z \approx \left[\frac{P(\nu_0)/S(\nu_0)}{4\pi(c/H_0)^2}\right]^{1/2}\left\{1 + \left(\frac{n}{2} + \frac{\Omega_0}{4}\right)\left[\frac{P(\nu_0)/S(\nu_0)}{4\pi(c/H_0)^2}\right]^{1/2}\right\}. \quad (3.6.17b)$$

The number of sources between Δ and $\Delta + d\Delta$ with flux densities between S and $S + dS$ is:

$$d^2N \sim n_0\Delta^2[1 + (2 - \Omega_0)\Delta]W(P, t)\left(\frac{\partial P}{\partial S}\right)_\Delta d\Delta dS, \quad (3.6.18a)$$

$$d^2N \sim n_0 z^2[1 - (2 + \Omega_0)z]W(P, t)\left(\frac{\partial P}{\partial S}\right)_z dz dS. \quad (3.6.18b)$$

For Δ and $t \approx -\Delta/H_0$ small, we may expand W and keep terms of lowest order only:

$$W(P, t) \approx W_0(P) - \frac{\Delta}{H_0}\left(\frac{\partial W}{\partial t}\right)_0. \quad (3.6.19)$$

We wish to find the number of sources with flux densities between S and $S + dS$; use (3.6.17) to get Δ as a function of P, write $d\Delta = (d\Delta/dP)dP$ and integrate over P:

$$dN \sim n_0 S^{-5/2} dS\left\{\int W_0(P)P^{3/2}dP + \frac{S^{-1/2}}{(4\pi)^{1/2}c/H_0} \times \left[(2n - 2)\int W_0 P^2 dP - \frac{1}{H_0}\int\left(\frac{\partial W}{\partial t}\right)_0 P^2 dP\right]\right\}. \quad (3.6.20)$$

By studying nearby galaxies we can determine $W_0(P)$. The spectral index n can be found from the source spectrum. The above result may be cast in two other useful forms. Earlier it was assumed that the evolution (i.e., the time dependence of W) does not depend on the appearance or disappearance of sources, but simply on the change in intrinsic power of the sources.

Suppose the time dependence of the luminosity of a particular source satisfies the following differential equation:

$$\frac{dP}{dt} = F(P, t) = Pf(P, t). \quad (3.6.21)$$

It then follows that

$$\frac{\partial W}{\partial t} = -\frac{\partial}{\partial P}(WF) . \tag{3.6.22}$$

This equation is of the form of a conservation law for sources moving along the P axis with "velocity" F. The last term in (3.6.20) may be rewritten as

$$-\frac{1}{H_0}\int\left(\frac{\partial W}{\partial t}\right)_0 P^2 dP = -\frac{2}{H_0}\int PF_0 W_0 dP . \tag{3.6.23}$$

Now, for the particular case of identical, synchronously evolving sources,

$$W(P, t) = \delta(P - P_0(t)) ; \quad \frac{dP_0}{dt} = F_0 = P_0 f_0 , \tag{3.6.24}$$

we find

$$dN \sim n_0 P_0^{3/2} S^{-5/2} dS\left\{1 + \left(\frac{P_0/S}{4\pi(c/H_0)^2}\right)^{1/2}\left[2n - 2 - \frac{2f_0}{H_0}\right]\right\} . \tag{3.6.25}$$

The quantity $f_0 = d(\ln P_0)/dt$ characterizes the rate of source evolution; f_0 has dimension s^{-1}. If sources were brighter in the past, then f_0 is negative. The square root quantity in (3.6.25) may be rewritten as

$$\frac{P_0/S}{4\pi(c/H_0)^2} = \frac{S_0}{S} , \tag{3.6.26}$$

where S_0 is the flux density from a source of luminosity P_0, located at rest in Euclidean space at the Hubble distance c/H_0.

In general, the weighted average quantities $\langle P\rangle^{1/2}$ and $\langle f\rangle$ may be defined as

$$\langle P\rangle^{1/2} \equiv \int W_0(P) P^2 dP / \int W_0(P) P^{3/2} dP , \tag{3.6.27a}$$

$$\langle f\rangle \equiv 1/2 \int\left(\frac{\partial W}{\partial t}\right)_0 P^2 dP / \int W_0(P) P^2 dP . \tag{3.6.27b}$$

As before, write $S_0 \equiv \langle P\rangle/[4\pi(c/H_0)^2]$, then

$$dN \sim n_0 \langle P\rangle^{3/2} S^{-5/2} dS\left\{1 + \left(\frac{S_0}{S}\right)^{1/2}\left[2n - 2 - \frac{2\langle f\rangle}{H_0}\right]\right\} . \tag{3.6.28}$$

This is the form in which theory and observations may be compared.

For unevolving sources at rest in Euclidean space, we would have

$$dN_0 \sim n_0 \langle P\rangle^{3/2} S^{-5/2} dS . \tag{3.6.29}$$

A comparison of (3.6.28) and (3.6.29) provides valuable information:

$$\frac{dN}{dN_0} = 1 + \left(\frac{S_0}{S}\right)^{1/2}\left[2n - 2 - \frac{2\langle f\rangle}{H_0}\right] . \tag{3.6.30}$$

From a plot of dN/dN_0 versus $S^{-1/2}$, the coefficient $2n - 2 - 2\langle f\rangle H_0^{-1}$ may be determined.

In the past, observers sought a comparison of the form: $dN \sim S^{-5/2-\mu}dS \sim d(S^{-3/2-\mu})$; that is, they searched for a deviation from the "3/2-power law." However, McVittie (1959*b*) showed that $\mu = 0$; indeed, (3.6.28) demonstrates this as well. For large S (i.e., small distances) the exponent is always 3/2 for the integral quantity N (5/2 for dN). The correction is of the form of an expansion in powers of $S^{-1/2}$, a quantity proportional to distance.

The most important qualitative consequence of the above result is that the correction to the "classical" formula (the coefficient of the $S^{-1/2}$ term) depends neither on the matter density nor on the deceleration parameter. This characteristic feature of the lowest order correction applies to observations of bolometric flux or to specific intensity of a power-law spectrum. It is well to remember that Ω_0 enters in the first order correction to the observed flux as well as in the expression for the distribution of sources with redshift. However, when these formulae are combined to yield the distribution of sources with respect to observed flux (with the redshift not measured), then there is an exact cancellation of the Ω_0 terms to lowest order. This peculiar result is of crucial importance since radio sources have power-law spectra and for optical sources the redshift is much more difficult to measure than the bolometric apparent magnitude.[2]

The measurements which make a comparison with (3.6.30) possible are certainly not without value. They are important for the verification of the fundamental assumption of the homogeneous spatial distribution of the sources and, especially, for the study of the evolution of the sources. Without independent data on the evolution of the sources, neither the deceleration parameter nor the average matter density can be found. This important point will be demonstrated in the next section.

3.7 On the Possibility of Choosing among Cosmological Models by Observations of Distant Objects

The quantities H_0 and Ω_0 which determine the cosmological model enter into the formulae connecting the intrinsic power of a source and the flux received at the Earth. These same quantities enter into the expressions for the number of sources with given redshift and/or with given flux. In these expressions, the number density of sources, expressed per unit comoving volume, will appear as a factor. For large z, the non-Euclidean character of space manifests itself; consequently, the effect of Ω_0 on the results is large. Moreover, if a large part of the matter is ionized gas, then the scattering of electromagnetic waves by free electrons can be important.

On the basis of observations, can one obtain information about cosmologically important quantities?

This is evidently possible if the intrinsic luminosity of a source is known

2. We note that in the review by Zel'dovich (1965*d*), formula (147) and subsequent formulae are in error.

of the number density of sources is known. Many authors have determined the relations needed for this analysis; see, for example, the preceding section of this chapter. However, the more difficult part of the problem is obtaining the necessary data about the sources.

Stars on the main sequence represent a single type of object for which physical theory yields conclusions regarding the connection between the spectrum and the intrinsic luminosity. For example, stars for which an empirical and theoretical connection has been established, between the period and luminosity, are the Cepheids. But Cepheids are not visible in distant galaxies. An observer naturally must consider distant objects with hope of his is to obtain reliable data. Now at large distances we essentially see galaxies (or quasars) as a whole, and we are not in a position to resolve separate objects or structural features in them. Concerning quasars, radio galaxies, or other objects visible at large distances, reliable methods for the determination of their intrinsic luminosities do not exist.

Nevertheless, in § 3.9 we will briefly discuss the best existing methods for the determination of the luminosity of ordinary galaxies and the methods for determining, with the help of that information, the parameters of the cosmological model. In this section we will discuss the fundamental question of the effect of source evolution on the determination of the parameters of a cosmological model. The remarks here will mainly be about the methods for the determination of the cosmological parameters through counts of radio sources and through the statistics of radio sources. For a long time, cosmologists placed great hopes in this method.

Can one determine the cosmological model when sources evolve? Below we show that, in the absence of a theory of the evolution, this is in principle impossible. Any model (i.e., any values of H_0 and Ω_0) can be combined with any observational data on the statistics of sources if the law describing the source evolution is in no way limited. Consequently, no data on the statistics of sources can, without knowledge of the evolution, allow one to determine H_0 and Ω_0. Inasmuch as one can consider H_0 known to an accuracy of $\sim\pm 50\%$, this proviso actually concerns the determination of Ω_0. Estimates of the uncertainty in the value of Ω_0 reach a factor of 100 ($0.01 \lesssim \Omega_0 < 5$). The differences in the accuracy with which H_0 and Ω_0 are known are explained below. An account of the scattering of light by electrons changes nothing. Any statement concerning the quantity Ω_0 based on a comparison of the statistical data with the formulae of relativistic cosmology is, in fact, based on some assumption concerning evolution; the reliability of the values obtained is no greater than the reliability of the assumptions about evolution.

Let us examine the question in its most general form. Let $\rho(P, t)$ be the density function for the sources; the number of sources with power between P and $P + dP$ per unit comoving volume is $dn = \rho(P, T)dP$. We utilize the letter P here for the specific luminosity in order to emphasize that we

are not necessarily concerned with the integral (bolometric) luminosity L. Relativistic cosmology gives us a connection between t and z [i.e., t $t(z, \Omega_0)$] and a connection between the flux received and the power emitted:

$$S = P\Phi(z, \Omega_0) \quad \text{or} \quad P = S\Phi^{-1} = S\Psi(z, \Omega_0) . \tag{3.7.1}$$

The influence of electron scattering can be included in the functions Φ and Ψ; if without scattering $\Phi = \Phi_0(z, \Omega_0)$, then in the presence of scattering $\Phi = \Phi_0 e^{-\tau}$, where τ is the optical depth. The comoving volume, in a layer in which the redshift changes from z to $z + dz$ and within a solid angle of 1 steradian is:

$$dV = \chi(z, \Omega_0)dz . \tag{3.7.2}$$

Now consider the number of sources in 1 steradian with flux between S and $S + dS$ and redshift between z and $z + dz$. Their power evidently lies between $P = S\Psi$ and $P + dP = (S + dS)\Psi$. Consequently,

$$\begin{aligned} d^2N \equiv N''(S, z)dSdz &= \rho(S\Psi, t)\Psi dS\chi dz \\ &= \rho[S\Psi(z, \Omega_0),\ t(z, \Omega_0)]\Psi(z, \Omega_0)\chi(\Omega_0, z)dSdz . \end{aligned} \tag{3.7.3}$$

Thus, for any function $N''(S, z)$ obtained from observations using any value of Ω_0, one can construct a corresponding evolutionary function

$$\rho(P, t) = \frac{N''[P\Phi(z, \Omega_0), t(z, \Omega_0)]}{\Psi[z(t), \Omega_0]\chi[z(t), \Omega_0]} . \tag{3.7.4}$$

Observations of optically unidentified radio sources (for them, z is unknown) yield a distribution with respect to the flux, i.e., a function $N'(S)$ satisfying

$$dN = N'(S)dS = dS\int_0^\infty N''(S, z)dz . \tag{3.7.5}$$

Evidently, these less complete observations can no better provide a basis for the determination of Ω_0. Statistical observations give information about the evolution of radio sources (see § 3.8) but not about cosmology in the narrow sense of the word.

Here it is necessary to outline the difference between the determination of H_0 and that of Ω_0 (for more detail, see § 3.9). Let us assume that measurements of several types of objects in our neighborhood give a definite value for their power P. We may also assume that this quantity evolves: $P = P(t)$. But it seems improbable that the evolution is at any singular point at the present time t t_0. Therefore the function $P(t)$ is given by the series expansion

$$P = P_0 + (t \quad t_0)P'_0 + \tfrac{1}{2}(t - t_0)^2P''_0 + \ldots , \tag{3.7.6}$$

which does not contain terms of negative or fractional degree; P_0 is assumed to be known, but in contrast P'_0, P''_0, etc., are not. Now a cosmological model connects $t - t_0$ and z (compare with § 3.5),

$$H_0(t_0 \quad t) \quad z - (1 + \Omega_0/4)z^2 , \tag{3.7.7}$$

so that

$$P = P_0 \left[z\left(\frac{P'_0}{H_0}\right) + z^2\left[\left(1 + \frac{\Omega_0}{4}\right)\left(\frac{P'_0}{H_0}\right) + \tfrac{1}{2}(P''_0/H_0^2)\right]\right]. \quad (3.7.8)$$

Now, for a computation of the observable flux S, we must remember (see 3.5) that

$$\frac{S}{P} = \frac{1}{4\pi D^2} \approx \frac{H_0^2}{4\pi c^2 z^2}\left[1 - \left(1 - \frac{\Omega_0}{2}\right) z\right], \quad (3.7.9)$$

so that

$$S \approx \frac{P_0 H_0^2}{4\pi c^2 z^2}\left[1 - \left(1 - \frac{\Omega_0}{2} + \frac{P'_0}{H_0 P_0}\right) z + \ldots\right]. \quad (3.7.10)$$

For a determination of H_0, only P_0 is necessary; however, the density parameter Ω_0 is tied to the poorly known quantity P'_0. The thorough research of Sandage (1972*a*, *b*, *c*) on the redshifts of ordinary galaxies gives $q_0 = 1.0 \pm 0.4$, which corresponds (for $p = 0$, $\Lambda = 0$) to $\Omega_0 \approx 2 \pm 1$. The true limits of error on this result are not clear. For more detail, see § 3.9.

The condition that t_0 is not a singular point in the evolution puts obvious limitations on the function $\rho(P, t)$ and, consequently, on the function $N''(S, z)$. We will not write them here.

If one assumes that no evolutionary effects occur, i.e., if one sets $\rho = \rho(P)$ with no time dependence, then one obtains definite functions $N''(S, z, \Omega_0)$ and $N'(S, \Omega_0)$ depending only on Ω_0. One can find the form of the function $\rho(P)$ through the study of nearby objects. Hopes did exist that one could find the quantity Ω_0 by this method. Obviously, objects with rather large power are necessary for this. Therefore, hopes at one time rested on radio galaxies and counts thereof. In reality, the observed function $N'(S)$ for radio sources was found to be incompatible with the time-independent function $\rho(P)$, regardless of the assumptions about Ω_0 (for any Ω_0, $0 < \Omega_0 < \infty$).

Attempts were made to specify the form of the evolutionary dependence in some kind of especially simple form; for example,

$$\rho(P, t) = \rho_0(P)(t/t_0)^{-\beta} \quad \text{or} \quad \rho[P, t(z)] = \rho_0(P)(1 + z)^{\beta}. \quad (3.7.11)$$

The total density of matter per unit physical (not comoving) volume is proportional to $\rho_m \sim (1 + z)^3$. Thus, the specific assumption about the radio galaxies means that their density (per unit physical volume) equals

$$\rho_{RG} = \rho[P, t(z)](1 + z)^3 \sim (1 + z)^{3+\beta} \sim \rho_m^{(1+\beta/3)}. \quad (3.7.12)$$

At the same time it is assumed that the form of their distribution with respect to power remains unchanged.

In addition, attempts were made to describe $\rho(P)$ at the present time by the simplest possible dependence consistent with the observations of the most nearby sources. After this, parameters characterizing the evolution were determined according to the observed integral curve; the sharp drop in this curve for small S (see the following section) was explained by the

scattering by electrons of the intergalactic plasma and was thus determined by the concentration of electrons and consequently by Ω_0. However, it is necessary to emphasize the difficulty of this approach; the form of the curve $N(S)$ and, in particular, its drop for small S can be caused not only by the effect of Ω_0 (through electron scattering) or by the deviation of the evolutionary curve from the simple power-law behavior $(1 + z)^\beta$, but also by evolution of the strength of the sources. In this case one can no longer describe $\rho(P, t)$ as the product of two functions $\rho(P)f(t)$ or $\rho(P)f(z)$. In the absence of a complete theory predicting the power-law dependence, there is no basis upon which to consider such dependence universal. Thus, we must emphasize that only the creation of a reliable theory for the evolution of sources will permit one to determine q_0, and consequently the type of model, from the observations. Let us once again stress that the value of q_0 can be found (under the assumption $\Lambda = 0$) not only from the m versus z or N versus z relations, but also, directly, by the discovery of all forms of matter in the very nearby surroundings of our galaxy (see subsequent discussions about this).

There is one important exception to the statement that any cosmological model can agree with any observed function $N(S)$ due to the arbitrariness introduced by evolution. This exception is the cosmological model of the stationary Universe (the steady state theory, to use the language of its authors), a model which is characterized in detail in § 24.10. In this theory one assumes the invariance of all properties of the Universe and of the matter filling it for the entire, infinite, time of its evolution. The theory of the stationary Universe excludes the possibility of evolution (on the average—with the exception of inevitable statistical fluctuations) and predicts a single, completely determined function $N(S)$. Its form depends (in practice, weakly) only on the function $W(P)$ describing the distribution of sources with respect to power, a function which can be determined by observations in our neighborhood.

Observations of radio sources give results that do not agree with the predictions of the theory of the stationary Universe. Indeed, sharp debate over the theory of the stationary Universe caused special attention to be directed toward the counts of radio sources—especially in England during the last two decades.

Below we will indicate other difficulties of this theory.

3.8 The Evolution of Radio Sources

Before turning attention to the methods for determining H_0 and Ω_0 by observations of ordinary galaxies, let us conclude the discussion of the information gathered by the counts of radio sources.

As was shown in the previous section, observational data give information about the evolution of radio sources. This information is incomplete inas-

much as the interpretation of these data depends on assumptions relating to cosmology (the values of H_0 and Ω_0). Keeping this in mind, we will present the results of one of the programs of research in this direction, the work of Longair (1966).

We note in passing that it was the outstanding, systematic research conducted by the Cambridge radio astronomers headed by Ryle that is the basis for this theoretical study.

Longair takes the flat model for interpreting the counts, i.e., the case of the critical density, $\Omega_0 = 1$. Then, $(1 + z) \sim t^{-2/3}$. He adopts different choices for the law of evolution. Satisfactory agreement with the observations down to $S = 0.4 \times 10^{-26}$ W (m^2 Hz)$^{-1}$ occurred for two variants. In the variant denoted *b* by Longair, it is assumed that the sources with absolute magnitude $M_r > -29.5$ (i.e., comparatively weak) do not evolve, while those with $M_r < -29.5$ evolve in such a way that their power grows as $t^{-2.2}$, i.e., in proportion to $(1 + z)^{3.3}$. The absolute magnitude M_r is related to the specific luminosity at 178 MHz by the relation $M_r = 33.95 - 2.5 \log_{10} P_{178}$. This law treats the region $z < 3$. For $z > 3$, it is assumed that there are no sources—either weak or strong—at all.

The second variant, denoted *d*, is based on the assumption that the density of sources with $M_r < -33$ is greater in the past, the density changing as $t^{-3.8} \sim (1 + z)^{5.7}$. Also, the density (per unit comoving volume) and the power of weaker sources are unchanging. For $z > 4$, there are no sources.

On the graph presented in Figure 14 appears the curve giving the depen-

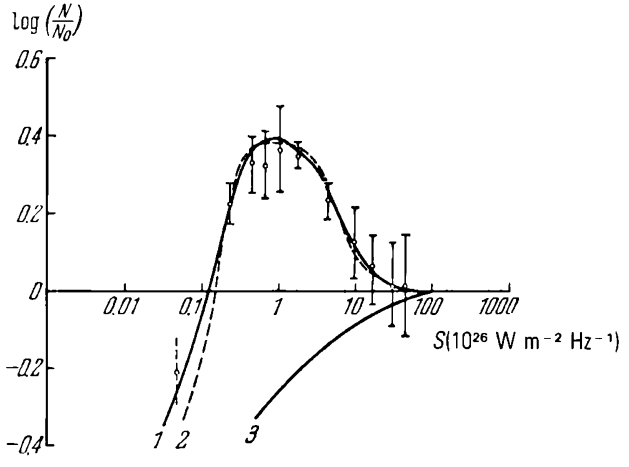


FIG. 14.—The ratio of the number of sources N with a flux greater than S to the quantity $N_0 = kS^{-3/2}$. The curve labeled 1 applies under assumption *b* (see text) for the evolution of sources; that labeled 2 applies under assumption *d;* while that labeled 3 applies when no account is made of evolution. The abscissa gives the flux at 178 MHz in units of 10^{-26} W m^{-2} Hz^{-1}.

dence on S of the ratio of $N(S)$ to $N_0 = AS^{-3/2}$; the curve is given on a logarithmic scale. The law $N_0 \sim S^{-3/2}$ corresponds to flat space without taking account of evolution and redshift, and the constant A is chosen in accordance with the number of nearby sources. Therefore, for large S, the ratio $N/N_0 \to 1$. For comparison purposes a curve is presented which does not account for evolution but which does account for cosmological effects in the case $\Omega_0 = 1$. This curve lies entirely below the abscissa. The rise in the curve in the region $10 > S > 1$ (10^{-26} W [m^2 Hz]$^{-1}$) is caused by the evolutionary effect—the growth in the number of sources or in their power with increasing z. The fall in the curve for $S < 1$ requires for its description an evolutionary effect of the opposite sign—the absence or sharp decrease

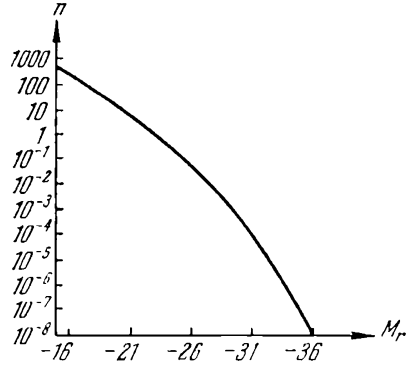


FIG. 15.—The distribution of radio sources located in region with $z < 0.01$ as a function of absolute radio magnitude M_r. The ordinate gives the number of sources n per steradian with $z < 0.01$ and having an absolute magnitude between M_r and $M_r + 0.5$.

in the number of radio galaxies for $z > 3$ or $z > 4$. The leftmost point was obtained by G. B. Sholomitsky in accordance with the data in the work of Kenderdine, Ryle, and Pooley (1966) and was not considered in the work of Longair; nevertheless, it falls right on the curve.

The graph given in Figure 15 presents the distribution of radio sources with respect to power. These sources are located in our vicinity ($z < 0.01$), so that any evolutionary effects are small for them.

The evolutionary assumptions of Longair correspond to a distinctive deformation of the curve for each different z. For variant b with $z = 3$ the curve appears in Figure 16; while for variant d with $z = 4$ the curve appears in Figure 17. The curve substantially changes its form!

It is especially necessary to emphasize the "cutoff" for $z = 3$ or $z = 4$. One must consequently assume that for $z > 3$ or $z > 4$, i.e., earlier than a definite time ($t \lesssim 5 \times 10^8$ yr), radio galaxies do not exist. Indeed, this estimate of the time for the initial, large-scale birth of radio galaxies coincides with estimates obtained from other considerations. One such estimate

based on the relation between the sizes of clusters of galaxies (with diameters of order 3 Mpc) and the average distance between them (of order 10 Mpc). For large clusters the diameters may be ~8 Mpc and the separation ~ 10 Mpc (see § 14.11). For $z \approx 3.5$ the clusters are contiguous and fill all of space. At the same value of z (Zwicky gives the value $z = 4.25$), consequently, the drifting apart of the clusters from one another occurs. Another estimate is connected with an examination of the heat balance of gas before the formation of galaxies (see the subsequent discussions).

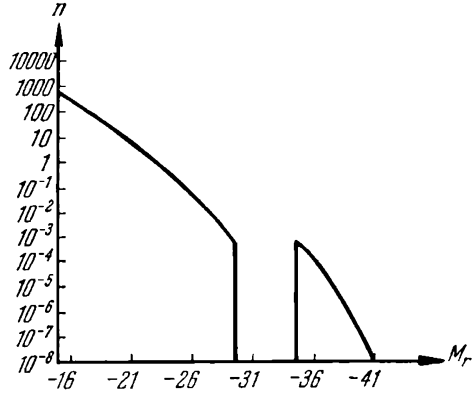


FIG. 16. The evolution of radio sources according to variant *b* (see text). The curve in the figure depicts the radio-source density at the time corresponding to a redshift $z = 4$. Only radio sources with absolute radio magnitudes $M_r < -29.5$ evolve. The density shown is number per unit comoving volume—with the comoving volume normalized to equal proper volume at the present time, and with that proper volume in the same units as were used in Fig. 15.

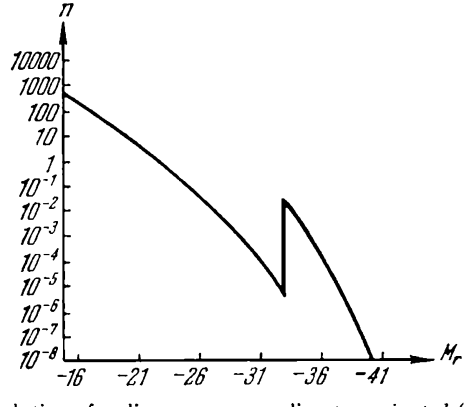


FIG. 17.—The evolution of radio sources according to variant *d* (see text). The curve $n = n(M_r)$ is shown for the time when $z = 4$. The number of sources with absolute radio magnitude $M_r < -33$ evolves. The density is given in the same units as are used in Figs. 15 and 16.

At first glance an unnaturally sharp change appears in the form of the curve $\rho(P)$ itself. As compared with the smooth behavior for the present, a discontinuity or hump appears in the past. This is shown in Figures 16 and 17. This means that present-day radio sources fall into two groups. According to Longair, the observational data are compatible with the assumption that the powerful, evolving group represents the quasars. What remains unclear is only why, at the present time, the two groups of sources give a smooth curve without kinks or discontinuities.

How unique is this picture of the evolution presented by Longair? If one assumes that all sources evolve similarly, i.e., that the entire curve $\rho(P)$ is displaced to the right or upward without a change in form, one cannot obtain the sharp drop in $N(S)/N_0(S)$ for small S. One cannot obtain it by choosing a value of z above which there are no sources, or by taking account of scattering.

Indeed, even if we do not see more distant sources as the detector sensitivity is increased (i.e., for smaller values of S), either because there are none beyond some definite limiting z_{MAX} or because they are hidden in a shroud of scattering, we will still see new, weaker sources. Only by means of a nonmonotonic curve of the form of that in Figures 16 and 17 can one achieve a halt in the fast growth of $N(S)$.

Longair did not calculate the scattering of light by electrons. Assume that ionized gas constitutes a fraction ξ of the average density. Then, using the relations

$$\tau = c\sigma \int n_e dl = c\sigma \left(\frac{6 \times 10^{23}}{\mu_e}\right) \int_0^z \rho_{\text{gas}}(z') \frac{dl}{dz'} dz' ,$$
$$\rho_{\text{gas}}(z) = \xi\rho_0(1 + z)^3 = \xi\Omega_0\rho_c(1 + z)^3 , \tag{3.8.1}$$

we obtain the following formulae for the scattering optical depth (we take $\mu_e = 1.16$ g):

$$\tau \approx 0.06\xi\Omega_0 h_0 \int_0^z \frac{(1 + z')}{(1 + \Omega_0 z')^{1/2}} dz' . \tag{3.8.2}$$

In the above, σ is the Thomson cross section $\sigma \approx \frac{2}{3} \times 10^{-24}$ cm^2. For $\xi = \Omega_0 = 1$, $h_0 = \frac{1}{2}$ we have:

$$\tau \approx 0.02[(1 + z)^{3/2} - 1] , \quad \text{and} \quad \tau \approx 1 \quad \text{for} \quad z \approx 13 . \tag{3.8.3}$$

For the evolutionary models of Longair, $\tau \lesssim 0.2$ at the cutoff value for the formation of sources (for $z = 4$). Consequently, it is only necessary to assume the existence of an insignificant supplementary change in the power of the sources to compensate for the effect of scattering. Remember that the evolutionary change in the power according to Longair's calculations varies as $(1 + z)^{3.3}$, i.e., for $z = 3$ the power becomes 100 times greater than at present.

The question of the statistical accuracy of the curve $N(S)$ is complicated

by the interdependence of two values $N(S_1)$ and $N(S_2)$. This occurs because $N(S_1)$ contributes to $N(S_2)$ for $S_1 > S_2$. In this connection Jauncey (1970) proposed that the differential curve dN/dS as a function of S be used instead. For practical purposes, it is advisable to find $\Delta_n N$, the number of sources in an interval of S, these intervals being determined by the geometrical progression $\Delta S_n = S(\alpha^n - \alpha^{n+1})$, with $\alpha < 1$. If $N(S)$ is a power-law function, one obtains a linear dependence of $\log \Delta_n N$ on n. Primarily, the differential or difference representation of the results of the counts of radio sources demonstrates that chance fluctuations in the number of nearby (weak) sources do not affect the subsequent course of the curve.

Objections were raised against the work of Longair *et al.* In particular, Kellermann (1972) maintained that the powerful sources in our neighborhood are actually not sufficient in number to present a noticeable evolutionary effect. In our opinion, however, the articles of Silk and Rees, and of Rees and Longair (1972) contain a convincing defense of the evolutionary picture. There are new data on $N(S)$ at 178 MHz and 5 GHz which confirm the general picture described above. See Longair (1974) for details.

Consider the relation between the counts of radio sources and the total background radiation. Especially interesting are the data on the total intensity of radiation at 178 MHz. Measurements show that this corresponds to radiation with an effective temperature of 80 K (in the direction of the galactic pole). Further data indicate that extragalactic sources have effective temperatures of 28 ± 5 K or, according to earlier data, 23 ± 5 K. In addition, one must subtract from the total, 3 K for RR. After considering all sources, including those which are unresolved, one finds a residual temperature between 15 K and 30 K.

An important feature of the background radiation is that scattering does not affect it. Scattering can change its direction, but not its frequency or intensity. Absorption, as distinct from scattering, of the radiation by the dilute intergalactic plasma is negligible even back to much higher redshifts 10^3. Thus, let us denote by P_1 the total radio power per unit comoving volume, with $P_1 = 4\pi \int P\rho(P)dP$. To account for evolutionary effects, $P_1 = P_1(t)$, but knowing the relation between t and z, we shall use the more convenient form $P_1 = P_1(z)$. The luminosity per unit physical volume equals $(1 + z)^3 P_1(z)$.

Now for a power-law spectrum ($\sim \nu^{-\alpha}$), the law describing the change of energy density per unit frequency interval at a given frequency as the Universe expands is $\epsilon_\nu \sim (1 + z)^{-3-\alpha}$. Then, using the relation (3.4.7),

$$H_0 dt = (1 + z)^{-2}(1 + \Omega_0 z)^{-1/2} dz ,$$

we obtain the expression

$$\epsilon_\nu = \frac{1}{H_0} \int P_1(z)(1 + z)^{-\alpha-2}(1 + \Omega_0 z)^{-1/2} dz , \qquad (3.8.4)$$

for the radiation density at the present time. It is convenient to pass from the radiation density to an effective temperature. For radio frequencies ($h\nu \ll kT$), these quantities are connected by the Rayleigh-Jeans formula:

$$\epsilon_\nu = \frac{8\pi kT}{\lambda^2 c}.$$

Thus, the general form for the temperature is

$$T = A \int P_1(z)(1+z)^{-\alpha-2}(1+\Omega_0 z)^{-1/2} dz\,. \tag{3.8.5}$$

If one isolates the evolutionary factor, i.e., if one writes $P_1 = P_0 f(z)$, where P_0 is the radio power at 178 MHz today, equation (3.8.5) assumes the form

$$T = 16 \int f(z)(1+z)^{-2-\alpha}(1+\Omega_0 z)^{-1/2} dz\,. \tag{3.8.6}$$

Let us assume power-law evolution from some value of z to $z = 0$. We further set $\Omega_0 = 1$. For $f = (1+z)^3$ and $\alpha = 1$, we then find $T = 32[(1+z)^{1/2} - 1]$ K. At $z = 4$, this would be $T \approx 40$ K. If one takes the more pronounced growth $f(z) = (1+z)^{4.5}$, which is necessary for a description of $N(S)$, then $T = 8[(1+z)^2 - 1]$ K ≈ 190 K for $z = 4$. Thus the assumption of general evolution of all sources, both weak and strong, with a cutoff at $z = 4$ is in conflict with the measured background temperature.

It is interesting to estimate the contribution to the temperature due to identified sources alone. One customarily characterizes them by the function $N(S)$, the number of sources per steradian for which the flux exceeds S. The number of sources with flux in the interval from S to $S + dS$ is $n(S) = -(dN/dS)dS$. Consequently, the total flux satisfies

$$g = \int_S^\infty S' \left(\frac{dN}{dS'}\right) dS' = SN(S) + \int_S^\infty N(S')dS'\,. \tag{3.8.7}$$

In the wide interval from $S = 0.3 \times 10^{-26}$ W m^{-2} Hz^{-1} to $S = 30 \times 10^{-26}$ W m^{-2} Hz^{-1} (where N is of order unity), the observed data are described by the power-law formula

$$N = 1.6 \times 10^3 S_{26}^{-1.7}\,, \tag{3.8.8}$$

where $S_{26} = S[10^{-26}\ \text{W m}^{-2}\ \text{Hz}^{-1}]^{-1}$. For such an exponent, the lower limit of integration $S_{26}^{\min}$ determines g, i.e., g fundamentally depends on the numerous weak sources. The flux turns out to be

$$g = 4 \times 10^3 (S_{26}^{\min})^{-0.7} \times 10^{-26}\ \text{W m}^{-2}\ \text{Hz}^{-1}\ \text{sr}^{-1}\,.$$

Expressed in terms of the effective temperature (for $\nu = 178$ MHz), the flux in these same units is

$$g = 960(T/K) \times 10^{-26}\ \text{W m}^{-2}\ \text{Hz}^{-1}\ \text{sr}^{-1}\,.$$

All sources with $S_{26} > 0.3$ give $T \sim 10$ K, in agreement with the results of Longair. Later data down to $S_{26} = 0.03$ gave a temperature which

significantly exceeds this value. As a consequence of this, the growth of $N(S)$ sharply slows for $S_{26} < 0.3$. However, if the decline in the growth of $N(S)$ were connected with scattering, then the sources, which would produce a flux $S_{26} > 0.03$ in the absence of scattering, would give $T = 44$ K. But such a temperature falls outside the permissible limits of 15 K and 30 K. Consequently, if one can in principle describe the decline in the growth of $N(S)$ as an effect of scattering, then one cannot describe the observed temperature without assuming that somewhere for $z > 4$ (or correspondingly $S_{26} \gtrsim 0.03$) there are no sources.

After the publication of Longair's work, subsequent measurements supported his predictions for small S.

The number of sources for S less than some definite value does not grow so quickly as formerly assumed. Longair assumed that the sharp cutoff $P_1(z) = 0$ occurs for $z > z_c$. As shown by Doroshkevich, Longair, and Zel'dovich (1970), one cannot exclude the possibility that $P_1(z) = \text{const.} = P_1(z_c)$ for $z > z_c$. Even a function of some other type, without any kind of cutoff, such as $P_1 \sim e^{-t/\tau}$, explains the measurements sufficiently well. The necessity of a cutoff is characteristic only of a power law $(1 + z)^{+\beta}$; such a law cannot be extrapolated indefinitely to $z = \infty$. There is moreover a physical reason to assume that no radio sources exist for large z ($z \gg 10$, for example). The reason is that the large radiation density hinders the acceleration of electrons. Remember that quasars with $z > 3.5$ are unknown. Here we state only that measurements of $N(S)$ do not give definite indications of the maximum value of z for radio sources.

Direct observations give a picture superficially reminiscent of an island Universe with our Galaxy at the center. The number density of sources at first grows with distance from the center, but then falls sharply at a definite distance. However, the aggregate of theoretical opinion interprets this apparent spatial inhomogeneity as an evolutionary effect in a spatially homogeneous world.

All of Longair's calculations are done with the assumptions $\Omega_0 = 1$ and $H_0 = 100$ km s^{-1} Mpc^{-1}. How sensitive are the calculations to these assumptions? The determination of $\rho(P)$ was based on a count of identified sources, with $z < 0.01$, for which the flux S is known. The scaling of this quantity ρ, a density, is expressed in terms of $\rho_{100}(P)$ (where $\rho \equiv \rho_{100}h_0^2$): $\rho(P)dP = h_0\rho_{100}(P_{100}/h_0^2)dP_{100}$. The scaling does not depend on Ω_0.

Although the form of the evolutionary curves $f(z)$ does not depend on H_0, it does depend on Ω_0. Alternatively, a given function $f(z)$ gives a similar result for any H_0, but to obtain identical observational results for different values of Ω_0, one must change the evolution law.

To explain how large these changes are, one can compare the data for the cases $\Omega_0 = \frac{1}{3}$, $\Omega_0 = 1$, and $\Omega_0 = 3$ when $z = 4$. It is easy to compute the quantities dt/dz, D^2, $D^2(dt/dz)$, τ, when $z = 4$. The consequent changes which occur on passing from one model to another are by no means small—

they reach an order of magnitude in size. However, it is necessary to compare them with the assumed evolutionary effects. In agreement with Longair's results, the evolutionary effects range in size from two to four orders of magnitude for the different models. In comparison with such evolution, the changes caused by different choices of Ω_0 are small. Thus, small changes in the characteristics of the evolution, leaving the picture intact, permit one to reconcile the observations with any cosmological model.

The conclusion to be drawn from this discussion is that observations attest to the evolution of radio sources but do not help in the determination of the parameters of the cosmological model. The next section is devoted to an explanation of the method for the determination of these parameters.

3.9 The Determination of H_0 and q_0 from Observations

The formulae presented in § 2.3–5 express various observable quantities in terms of the parameters of a cosmological model. Through a comparison with observations these formulae must in principle lead to a determination of the two parameters H_0 and q_0 (or Ω_0) and thus to a knowledge of all the parameters of a homogeneous, isotropic cosmological model. As was emphasized in the Introduction, such a model very accurately describes the present state of the Universe. This belief is supported by observations, the first and foremost of which are the observations of RR. The RR confirms the qualitative picture of the isotropic expansion of the Universe but gives no clue whatsoever to the values of H_0 or Ω_0. This radiation will be discussed in detail later.

Of all the aforementioned methods of testing cosmological theory through observations of distant objects, the most suitable for the determination of H_0 and q_0 is based on the relation between the apparent magnitude and the redshift: m versus $\log z$. All other methods, (for example, counts of galaxies or radio sources; see Hewish, Readhead, and Duffet-Smith 1974) are fraught with great difficulties as well as with the errors of observations, selection effects, and so forth. The relation of m to $\log z$ appears as formula (3.4.21), and the expansion of this formula for $z < 0.3$ as equation (3.5.25). Let us rewrite this formula once again:

$$m_{\text{bol}} = 5 \log \left\{ \frac{1}{q_0^2} \left(q_0 z + (q_0 - 1)[(1 + 2q_0 z)^{1/2} - 1] \right) \right\} + C_1 ; \quad (3.9.1)$$

$$m_{\text{bol}} \approx 5 \log z + 1.086(1 - q_0)z + C_1 , \qquad z < 0.3 , \quad (3.9.2a)$$

where

$$C_1 = M_{\text{bol}} - 45.06 - 5 \log H_0 . \quad (3.9.2b)$$

Formulae (3.9.1) and (3.9.2), with H_0 expressed in s^{-1}, show that one need only measure m_{bol} for two sources with *known* M_{bol} to determine q_0 and H_0. But all of the difficulties of the cosmological problem converge on

of the determination of the quantity M_{bol} for any bright source seen from afar. In the absence of a "standard candle," astronomers use the potentially large number of sources whose luminosity is only known approximately and which are located at various distances from us.

To be specific, one assumes (an assumption verified by observations; see below) that in diverse clusters of galaxies, the brightest cluster members (or one member of each cluster) all have equal absolute magnitudes. One then constructs a diagram of apparent magnitude (m) versus the logarithm of the redshift ($\log z$) for such objects.

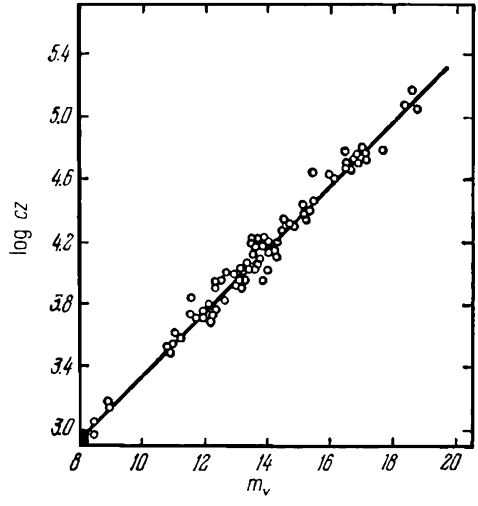


FIG. 18.—The redshift–apparent-magnitude (z–m_v) diagram for brightest cluster galaxies according to Sandage (1972*b*). On the ordinate is plotted $\log_{10} cz$ with c the speed of light in km s^{-1}; on the abscissa is m_v, apparent magnitude in the visual band.

In Figure 18 appears the Hubble diagram of redshift versus apparent magnitude for the brightest cluster galaxies. This graph is based on data from the study by Sandage (1972*a*, *b*, *c*). A similar diagram permitted Hubble to establish the redshift law itself in 1929. At that time Hubble had observational data available which roughly covers the region corresponding to the black rectangle in the lower left corner of the diagram.

From then to the present, observational capabilities have increased significantly. Indeed, not long ago it seemed that progress would be quite great. The hopes rested on the existence of powerful radio sources which, with the help of modern radio telescopes, are visible at larger distances than are ordinary galaxies. However, as was shown in § 3.7 and § 3.8, the strong evolution of these sources makes them very unsuitable for an attack on the cosmological problem.

During the last decade or so, quasars were discovered. In optical power

these objects exceed galaxies by one and a half to two orders of magnitude. However, the quasars are also unsuitable for the cosmological problem. In distinction to the brightest cluster galaxies, which are of about the same absolute magnitude M, the quasars have a large spread in absolute magnitudes. In addition, they too manifest a significant evolutionary effect. The diagram of m versus log z for them is depicted in Figure 19 and demonstrates no definite relation. Some studies have divided quasars into separate classes and studied the m versus log z relation for each class. For example, see Setti and Woltjer (1973), Bahcall and Hills (1973), and Burbidge and O'Dell (1973). So let us return to the m versus log z diagram for galaxies.

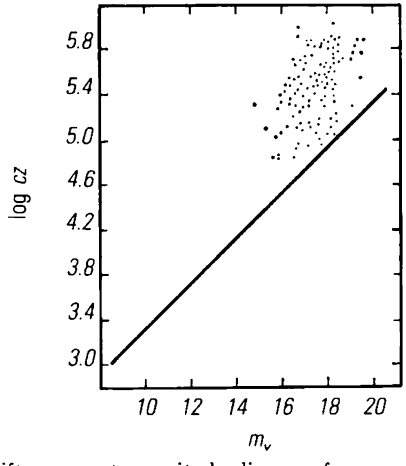


FIG. 19.—The redshift–apparent-magnitude diagram for quasars. The units are the same as in Fig. 18.

Before proceeding further, though, notice that what is plotted is not m_{bol}, but rather the so-called photovisual magnitude m_v, i.e., a quantity obtained by selectively filtering the light. It thus represents not the total flux (as does m_{bol}) but only that from a definite part of the spectrum. A correction is introduced in m_v because a selective receiver measures the intensity emitted for large redshift sources in different parts of the spectrum from those at small redshifts. The so-called K-correction then normalizes all measurements to the same part of the spectrum. Of course, the intrinsic spectra are considered known from studies of nearby galaxies. In addition, the apparent magnitudes are corrected for interstellar absorption in our Galaxy. Other corrections are also introduced into the apparent magnitudes of galaxies, but we will not discuss them here. The reader can find a review of this question and also a list of references in the article by Peach (1972). After accounting for the indicated corrections, naturally the formulae for photovisual magnitudes have the same form as do formulae (3.9.1) and

(3.9.2). We will make use of these formulae, keeping in mind the caveats regarding astronomical corrections.

The points on the diagram of Figure 18 lie, with little scatter, along a line that is possibly somewhat curved in its upper part, but that is otherwise straight. Formulae (3.9.1) and (3.9.2) with M_{bol} = const. describe this relation for galaxies quite well. The term 5 log z describes the straight line in the diagram; this term is important for small z. The second term, $1.086(1 - q_0)$, is small in comparison to the first and describes the slight curving of the relation for large z. The diagram is strong evidence that the initial assumptions we made are correct, at least in the first approximation. We recall these initial assumptions:

1. The Universe is expanding, with galaxies receding from us at velocities proportional to their distances (linearity of the m versus log z relation).

2. The expansion is isotropic (galaxies located in different regions of the sky all lie on the same curve in the diagram).

3. The absolute magnitudes of the brightest cluster galaxies are roughly equal. If they were not, a linear relation would not have been obtained; a scatter would have been obtained as on the diagram for quasars.

4. The evolutionary effect is not very large. If it were, the relation could be markedly curved (see below for detail).

In order to determine H_0 from observations, one must know the absolute magnitude of some (one) galaxy. From the linear (small z) part of the diagram, we can find H_0. The linear part of the diagram includes all the components of formula (3.9.1) except for the term $1.086(1 - q_0)z$, which is small when z is small. This term describes the small bending of the curve for large z and can lead to a determination of q_0.

For a solution of the first problem—finding H_0—M is found for the brightest galaxies of nearby clusters, for example, the cluster in the constellation Virgo. We will briefly describe here the procedure for making this determination, referring for details to the review by Sandage (1962). New data are presented in the articles in the proceedings of IAU Symposium 44 (Evans 1972) and in the literature cited therein.

The method consists of the sequential determination of the distance to more and more remote objects until the Virgo cluster is reached. First one finds the absolute magnitudes of stars in the nearest open star clusters. In practice, one uses the Hyades cluster.

The distance to the Hyades cluster is found by the following method. A measurement of the proper motions of individual stars in the cluster due to their motion along with the cluster permits a determination of the angle α between the direction from us to the cluster and the direction of its motion in space relative to that of the solar system. This angle equals the angular distance on the sky between the cluster and the point on the sky where the extension of the directions of motion of the separate stars intersect. Then, using the radial velocities of the stars, v_{rad}, one determines the

displacement velocity v_t of the stars in the plane defined by the motion of the solar system and of the cluster: $v_t = v_{rad} \tan \alpha$. Relating this velocity to the average angular velocity displacement of the cluster β, one finds the distance to the cluster as $R = v_t/\beta$. For the Hyades cluster, the distance is about 40 pc (1.2×10^{19} cm). Then, knowing the distance and the apparent magnitude of the stars, one can determine the absolute magnitude for all stars in the cluster, and in particular, for those which are in the hydrogen-burning stages (stars on the "main sequence" of the color-magnitude diagram). The absolute magnitudes of stars on the main sequence depend only on their color (i.e., on the spectral class) and, after a determination of the distance to Hyades, can be considered known.

The next step consists of the determination of distances to other, more distant clusters. This is done by a comparison of the apparent magnitudes of stars on the main sequence in these clusters with those in the Hyades cluster. In this way one can determine the distance to all clusters in our Galaxy.

Knowledge of these distances permits one to find the absolute magnitude of several periodic variable stars—the Cepheids—found in these clusters. The variability of the Cepheids obeys a remarkable relation that connects the period τ_{Ceph} with the luminosity (or absolute magnitude M_{Ceph}). Knowing the absolute magnitude of several Cepheids in a cluster, one finds the "zero point" of this relation; i.e., one finds the M_{Ceph} which corresponds to a given τ_{Ceph}. Individual bright Cepheids are visible at large distances. In particular, they are visible in one of the nearest spiral galaxies, M31 (Andromeda). By measuring the period of Cepheids in M31, their absolute magnitude is obtained through the τ_{Ceph}–M_{Ceph} relation and by comparing M_{Ceph} with m_{Ceph}, the distance may be found to M31. Unfortunately, Cepheids are not luminous enough to be visible in the more distant galaxies of the Virgo cluster.

Now we must take still another step. Many globular star clusters are visible in the galaxy M31. Hence, knowledge of the distance to M31 permits one to calculate the absolute magnitude M_{glob} for the globular clusters; ($M_{glob} \approx -10$). Now globular clusters can be seen in the galaxy M87 that is part of the Virgo cluster. By measuring the apparent magnitude of the globular clusters in M87 and knowing the globular cluster absolute magnitudes M_{glob}, one determines the distance to the Virgo cluster (about 1.7×10^7 pc). Remember that, for distances not too large on the cosmological scale, the meaning of distance is unique.

Finally, from this distance and the apparent magnitude of the brightest galaxy in the Virgo cluster (NGC 4472), one finds the absolute magnitude M_{gal} of this brightest galaxy; it satisfies $M_{gal} = -22.5$.

Instead of the aforementioned method of using Cepheids or globular clusters as distance indicators, one also uses other bright objects with more or less certain, known absolute magnitudes. Examples of these are the

brightest visible stars, novae, and supernovae. In addition, the linear sizes of regions of ionized hydrogen (H II regions) are used. The size of the brightest of these H II regions appears rather constant and therefore can also serve as a distance indicator.

It is evident that mistakes are possible at each step of the long progression just described. As a result, H_0 can be considered known at present with an accuracy hardly better than a few tens of a percent and perhaps with an accuracy of not higher than $\pm 50\%$. Recall that the determination by Hubble in 1936 gave for H_0 the value 530 km s^{-1} Mpc^{-1}. The modern determinations by Sandage, de Vaucouleurs, van den Bergh, and others give values between 100 km s^{-1} Mpc^{-1} and 50 km s^{-1} Mpc^{-1}, or even smaller still. Reviews appear in the collection edited by Evans (1972). For new data see Sandage (1972*a*, *b*, *c*, 1973*a*, *b*) and Sandage and Tammann (1974*a*, *b*, *c*, 1975*a*, *b*). In the subsequent development we must remember this indeterminacy. To allow for this uncertainty, in subsequent numerical estimates we will use: $H_0 = 100h_0$ km s^{-1} Mpc^{-1}.

At first glance the general tendency has a quite "threatening" character. From the time of the first determinations of H_0, its value, as derived from more and more precise observations, has (until recently) only decreased with the passage of time.

The general reason for this "decrease" is related on the one hand to confusing brighter objects in other galaxies with weaker ones (for example, in mistaking luminous gas clouds for stars). In addition, it is related to our realization that at a large distance, for which the number of objects is large, we selectively choose the brighter objects (see Tammann 1973*a*).

Much interest therefore exists in obtaining lower limits to H_0. In this respect, Pikel'ner (1974) noticed that the existence of radioactive isotopes permits one to determine the time since the epoch of active nucleosynthesis in the Galaxy. In an article by Fowler (1972), the most modern value is given for this time, namely $(11.7 \pm 2) \times 10^9$ yr. For a review see Schramm (1974). We shall assume that this quantity is also of the same order for other galaxies. Now in the period of active nucleosynthesis, galaxies must be many times brighter than now (Partridge and Peebles 1967*a*, *b*). Up to now, however, we have not seen such superbright galaxies. Hence their age of 12×10^9 yr must correspond to a redshift $z \gtrsim 1$; otherwise such objects would be seen. But, for this, it is necessary that $H_0 \gtrsim 40$–50 km s^{-1} Mpc^{-1}.

By quite rough reasoning then, Pikel'ner obtains a valuable lower limit to H_0 that is in no way connected with the long chains of reasoning necessary for a determination of the distance scale. An analysis of the question of the age of globular clusters, the determination of which is based on the theory of the evolution of main-sequence stars, leads to an analogous conclusion. Iben and Rood (1970) give $t_0 = (10 \pm 3) \times 10^9$ yr for this age, while

Böhm-Vitense and Szkody (1973) find about 1.4×10^9 yr; see also Talbot and Arnett (1973*a*, *b*).

Now let us proceed to the determination of q_0. As is apparent from formula (3.9.1) or (3.9.2), q_0 describes the nonlinearity in the relation between m and $\log z$. The quantity q_0 is determined directly from the diagram and does not require knowledge of the extragalactic distance scale as is necessary for the determination of H_0. Nevertheless, as we shall see later, the precision in the determination of q_0 is very low. In accordance with the assessment by Sandage (1972*a*, *b*, *c*) based on an interpretation of the data in Figure 18,

$$q_0 = +0.96 \pm 0.4 \quad \text{(probable error)} . \tag{3.9.3}$$

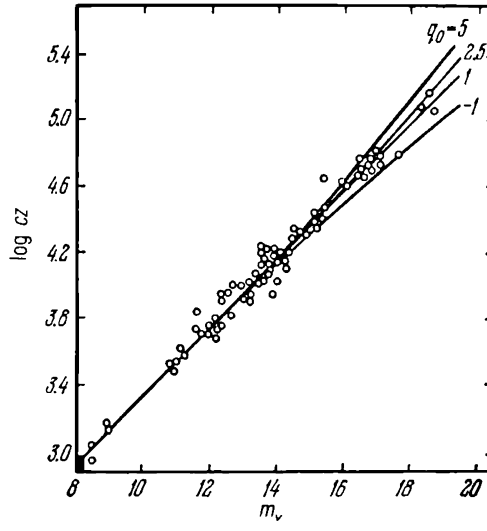


FIG. 20.—The theoretical redshift–apparent-magnitude curves for various values of q_0

On Figure 20 lines are shown which correspond to the relation for various values of q_0. Thus, q_0 is not known with great certainty. In addition, we must add the following remark. As we have repeatedly mentioned, the relation between m and $\log z$ is curved not only due to the dependence on q_0, but also as a result of the evolutionary effect, i.e., because the absolute magnitude M in (3.9.2b) cannot be considered constant in time.

Let us rewrite formula (3.9.1) in a form which takes account of evolution. For this we use the expression (3.7.10). Passing to the notation of magnitudes, we find

$$m_{\text{bol}} \approx 5 \log z + 1.086 \left(1 - q_0 - \frac{0.92}{H_0} \frac{dM_{\text{bol}}}{dt}\right) z + C_1 . \tag{3.9.4}$$

Thus, the observed quantity (3.9.3) is in essence the sum of the second and third terms in the bracketed expression above. Theoretical estimates of the

olution of galaxies (see Sandage 1970; Spinrad and Taylor 1971; Tinsley 1968, 1972, 1973) show that the third term can be just as important as the term. The quantity dM_{bol}/dt can be taken very roughly to be

$$dM_{bol}/dt \approx (-0.05 \pm 0.02) \times 10^{-9} \text{ mag yr}^{-1}.$$

At the present time, therefore, one cannot speak confidently about a definitive determination of q_0. Problems surrounding these determinations can be further aggravated by observational selection effects (see Peach 1972).

We can, however, draw one important conclusion. In accordance with the "steady state" cosmological theory, which at one time was defended by Hoyle and several other astronomers, the Universe is unchanging in time. This leads to $q_0 = -1$ and $dM_{bol}/dt = 0$. Figure 20 presents the curve corresponding to $q_0 = -1$. As Sandage emphasizes, it is extremely unlikely that this curve agrees with the observations. In this connection, recall that the evidence for the evolution of radio galaxies, presented in 6.8, also refutes the "steady state" theory.

6.10 Observable Quantities in a Universe That Is Homogeneous Only on the Average

The Universe is homogeneous over scales sufficiently large that many galaxies are included. However, the inhomogeneity of the matter distribution over small scales is evident. We are indebted to this inhomogeneity for the existence of stars, galaxies, and clusters of galaxies. Part III of this book is devoted to a detailed study of the influence of inhomogeneity on the dynamics (geometry and motion) of the Universe and on the propagation of light. However, it is advisable to discuss here one particular question concerning the general problem of the effects of inhomogeneity. This is the question of the effect of a large-amplitude inhomogeneity of small spatial extent l; i.e., a local density inhomogeneity with a local density significantly different from the average density.

Why is it that we can and should examine this particular question in that part of the book dedicated the theory of a strictly homogeneous Universe? The main reason is that, in practice, an inhomogeneity of small scale influences neither the dynamics of the Universe as a whole nor the redshift of distant objects.

Very roughly, one can suppose that the law for the change in the separation between two objects located some distance R from each other depends on the mass of the matter in a sphere of diameter R. If $R \gg l$, then many "inhomogeneities," roughly $N \approx (R/l)^3$, are found in the sphere. Moreover, the deviation of the density of this sphere from the average density is small. A rough estimate for the fluctuations in the number of random inhomogeneities is $\Delta N/N \sim N^{-1/2}$. In Part III we will see that the deviation

of the average density can be even smaller. Hence the inhomogeneities do not influence the large-scale dynamics.

However, it is necessary to follow in some detail the propagation of a bundle of rays which carry energy from a distant object (a galaxy or a quasar) to an Earthbound observer. Those inhomogeneities which specifically influence the propagation of light rays are those located near the paths of the rays. Specifically this means inhomogeneities in the matter located within the cone of the rays received by the observer. The volume of this cone is negligible in comparison with the volume of the sphere of diameter R. Therefore, an inhomogeneity which one can disregard in describing the dynamics of the Universe can be essential for a correct description of many observable quantities. Moreover, consideration of the inhomogeneous case permits one to understand more deeply the physical meaning of what occurs in a strictly homogeneous Universe.

In this section, then, we will study the effect of inhomogeneity in the Universe on the angular size and apparent magnitude of distant objects. For a homogeneous Universe, the relations of these quantities to distance, or, more precisely, redshift, appear in the formulae already derived. But for the application of these formulae to be valid, the homogeneity must satisfy more restrictive requirements than if simply the expansion law for the homogeneous model is to be valid.

The account in this section is based on several Soviet works: those of Zel'dovich (1964*b*), Dashevsky and Zel'dovich (1964), and Dashevsky and Slysh (1965). At the time of the symposium in Byurakan (1966), American astronomers noted that analogous remarks had been made by the physicist R. Feynman. Not long ago Kantowski (1969) and Dyer and Roeder (1973) deduced similar formulae.

Recall the remarkable fact that the angle subtended by a distant object passes through a minimum as the distance to it increases. As the horizon is approached, the angle increases without bound (see § 3.3). A vivid explanation of this follows from an examination of the gravitational focusing of the light rays emanating from the object AB and coming together at the point of observation (see Fig. 21). The bending of the light rays is the same phenomenon as the famous deviation of the light rays by 1″.75 near the disk of the Sun. That the bending of the rays OA and OB increases the angle between the rays at point O for a fixed size AB is readily apparent. Thus, the effect of the matter within the cone is to displace the ray AO downward (along the length of the ray, from right to left) and the ray BO upward. Matter located outside the cone OAB, if distributed in an axisymmetric fashion, does not affect the angle θ, independent of the amount of such matter. An inhomogeneous distribution of matter outside the cone of light rays also influences observable quantities. Calculations of the effect were done by Penrose (1966), Bertotti (1966), and Gunn (1967).

The condition for the applicability of the homogeneous-Universe formulae

For example, it follows from the criterion stated above: The density within the cone of light rays must not differ from the average density. If intergalactic gas (dust, and/or neutrinos constitute a large part of the density, then this condition can be fulfilled. But consider the opposite case wherein all of the matter is condensed into separate galaxies (Fig. 21). Then a given galaxy can fall in the cone either wholly (type E_1), partially (type E_2), or not at all. Moreover, it is possible that not even one galaxy lies within the cone. Indeed, in practice we often observe objects which are not hidden by other galaxies along the line of sight. If within the sphere whose center is at

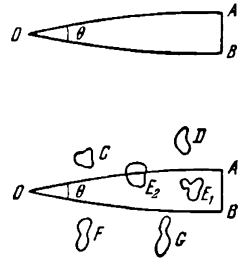


Fig. 21. The measurement of the angular size θ of an object AB in a homogeneous Universe (*above*) and in a Universe consisting of isolated clusters of material (*below*).

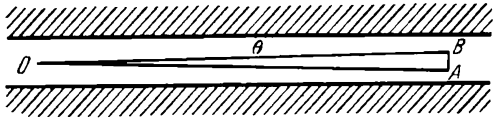


Fig. 22. A sketch elucidating the calculation of the angular size of an object in a Universe that is homogeneous only on the average. The matter density in the unshaded region is zero.

O and in which the object AB lies, there is a huge number of galaxies, the deviation of this number from the average is insignificant. However, if the object AB is small, then the volume of the cone is small and, corresponding-ly, the probability is large that the density in the cone differs from the average density. In particular, it is possible that the cone is empty.

One can successfully calculate the angle for the case of an empty cone in the following way (Zel'dovich 1964*b*). Consider space filled with matter of density $\rho = \langle\rho(t)\rangle$ outside a cylinder (Fig. 22) and empty inside the cylinder. This space can represent a homogeneous Universe on which is superposed a density perturbation $\delta\rho = -\langle\rho(t)\rangle$ inside the cylinder. For a homogeneous Universe, the problem of the propagation of the rays is exactly soluble, even if the Universe is nonstatic. Since the perturbed region is small, one can apply the Newtonian approximation to the gravitational potential theorem. Inasmuch as the angle θ is small, the radius of the cylinder is small

in comparison to its length and, locally, the field possesses cylindrical symmetry. In order to apply perturbation theory, the deflection angle of the ray must be small in absolute value, $\Delta\theta \ll 1$, but not necessarily small with respect to the homogeneous-Universe angle θ. Referring for details of the calculation to the article cited above, we present the result for the critical-density case, when $\rho = \rho_c$. (The deflections for the cases $\Omega_0 < 1$ and $\Omega_0 > 1$ were calculated by Dashevsky and Zel'dovich 1964). Under these circumstances, $\Omega_0 = 1$, the relation between the angle and the redshift in a completely homogeneous Universe is given by the expression

$$\theta = \frac{lH_0}{c}[2(1-\Delta)(1-[1-\Delta]^{1/2})]^{-1} \equiv \frac{lH_0}{cf(\Delta)}. \qquad (3.10.1)$$

In the case of an empty cone this becomes

$$\theta' = \frac{lH_0}{c}\left[\frac{2}{5}(1-[1-\Delta]^{5/2})\right]^{-1} \equiv \frac{lH_0}{cf_1(\Delta)}. \qquad (3.10.2)$$

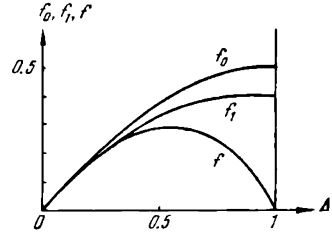


FIG. 23.—The dimensionless angular-diameter distance as a function of the redshift parameter Δ. The curve labeled f_0 applies to an empty Universe; that labeled f applies to a homogeneous Universe with an average density satisfying $\Omega = 1$; and that labeled f_1 applies to an inhomogeneous Universe with a density satisfying $\Omega = 1$ on the average but equal to zero within the cone of light rays from the source to the observer. The formulae for f_0, f, and f_1 are $f_0 = \Delta(1 - \frac{1}{2}\Delta), f = 2(1-\Delta)[1-(1-\Delta)^{1/2}], f_1 = \frac{2}{5}[1-(1-\Delta)^{5/2}]$.

The curves $f(\Delta)$ and $f_1(\Delta)$ appear in Figure 23. On the same graph, for comparison, is the curve of the function $f_0(\Delta)$ corresponding to the case of a completely homogeneous model with $\rho = 0$ everywhere—both inside the cone and outside (the Milne model). The curve f_1 does not have a maximum. This confirms the qualitative considerations regarding the physical reason for the minimum in the angle θ in a completely homogeneous Universe.

On the other hand (in the case where matter is concentrated in galaxies), one must consider the curve for θ as an average. This is so because for objects seen through an empty light-ray cone, the θ' curve gives the angle. But for other objects seen through a cone in which there are galaxies (such as E_1 and E_2 in Fig. 21), the angle is much greater than θ' because the galaxies act as gravitational lenses. This result is in agreement with the findings of Zwicky (1959; see also Klimov 1963).

In reality, there is some matter between galaxies. This is primarily intergalactic gas, whose density is presumably about 10^{-29} to 10^{-32} g cm^{-3}. This gas is probably distributed more or less uniformly. If this is so, then there is always gas in the cone of light rays. By the same method as above, it is easy to obtain the formulae for this case too (Dashevsky and Slysh 1965). Let the density, ρ_{gas}, of the uniformly distributed intergalactic medium constitute a fraction γ of the total density ρ, so that $\rho_{gas} = \gamma\rho$. The pressure is negligible, $P = 0$, and we again assume $\Omega_0 = 1$. Then, as calculations show,

$$\frac{\Delta}{\alpha}(1-\Delta)^{(5-\alpha)/4}[1-(1-\Delta)^{\alpha/2}]\,; \quad \alpha = (25-24\gamma)^{1/2}\,. \tag{3.10.3}$$

This expression has a maximum for any $\gamma \neq 0$ and tends to zero at the horizon. The smaller the fraction γ, the closer the maximum lies to the horizon.

Notice, finally, that everything said refers also to the calculation of the amount of light received by an observer from a distant object. The amount of light equals the product of the brightness and the solid angle. The brightness of the object depends on the properties of the object itself and on the redshift, while the solid angle is proportional to θ^2. Therefore, in the case of an object within an empty cone, a correction factor $(\theta'/\theta)^2$ enters into the expression for the apparent magnitude. Such a correction is also necessary in that case when both θ and θ' are themselves so small that the solid angle is not resolved.

3.11 The Kinetic Equations for Photons

For the calculation of observable quantities, we have so far considered the propagation of light from separate sources to an observer. For many purposes, however, it is important to know the average density of radiation in space from all sources, as well as the spectral distribution of the radiation — in other words, the average density of radiation far from discrete sources.

One can approach the question of the average density of radiation due to discrete sources by a method completely different from that of considering the propagation of light from sources to an observer.

In a homogeneous and isotropic Universe, the radiation must be homogeneous and isotropic on the average. Consequently, one can characterize it by a function of two variables, the frequency ω and the time t. The number of photons in a given frequency interval $d\omega$ per unit volume at time t is

$$dn_\gamma = \phi(\omega, t)d\omega\,. \tag{3.11.1}$$

Correspondingly, the energy density of photons in the interval $d\omega$ is

$$d\epsilon_\gamma = \hbar\omega dn_\gamma = \hbar\omega\phi(\omega, t)d\omega \equiv \xi(\omega, t)d\omega\,. \tag{3.11.2}$$

In an isotropic radiation field, an exchange of photons between neighboring volume elements changes nothing because the spectra and density are everywhere the same. This makes a local approach to the calculation of the function ϕ possible.

Let us thus form the partial differential equation for $\phi(\omega, t)$. This equation is a particular case of the kinetic equation for a distribution of particles with respect to coordinates and to velocity or momentum. Such an equation is very complicated in the general case in which are included the effects of the curvature of spacetime, as well as those of the scattering, absorption, and emission of the particles under consideration. However, the equation and conclusions derived therefrom become simple for the particular case in which the metric of spacetime corresponds to a homogeneous and isotropic model and in which the particles (in this case photons) are also distributed homogeneously and isotropically at every point in space. From general considerations it is clear that if at some instant the distribution possesses these properties, then the distribution will possess them at all later times. Hence, for all time one need only consider the function of two variables $\phi(\omega, t)$ instead of a more complicated function of the seven variables $\boldsymbol{r}$, $\boldsymbol{p}$, t, or r, n, ω, and t. The position vector $\boldsymbol{r}$ and the momentum vector $\boldsymbol{p}$ each contain three variables. But one may use instead $\boldsymbol{n} = (c/\hbar\omega)\boldsymbol{p}$, where $\boldsymbol{n}$, a unit vector giving a direction, corresponds to only two variables. It is further obvious that in the isotropic case, inelastic photon scattering does not enter into the equation. At first we will also neglect the emission and absorption of photons and will take account only of the redshift.

Thus let us examine that group of dN_γ photons which is located in the comoving volume V in the interval of frequency $d\omega$:

$$dN_\gamma = V\phi(\omega, t)d\omega\,.$$

In the course of the expansion, the number dN_γ remains unchanged if neither absorption nor emission occurs. We must, however, account for the changes in V and ω. Thus, by the general rules of differential calculus,

$$\frac{d}{dt}(dN_\gamma) = \frac{d}{dt}[V(t)\phi(\omega, t)d\omega(t)] = 0\,. \tag{3.11.3}$$

Recall that

$$\frac{dV}{dt} = 3HV\,, \quad \frac{d\omega}{dt} = -H\omega\,, \tag{3.11.4}$$

so that we obtain

$$3HdN_\gamma - HdN_\gamma + dN_\gamma\left(\frac{1}{\phi}\frac{d\phi}{dt}\right) = 0\,,$$

or

$$2H\phi + \frac{\partial\phi}{\partial t} - H\omega\frac{\partial\phi}{\partial\omega} = 0\,. \tag{3.11.5}$$

Introduction of the occupation number of photons, n_{occ}, simplifies this equation. The occupation number is defined as the ratio of the number of photons in a given element of volume and frequency interval (and solid angle in the general case), dN_γ, to the number of separate normal modes of oscillation of the electromagnetic field, $d\mu$,

$$d\mu = \frac{2}{(2\pi\hbar)^3} V dp_x dp_y dp_z = \frac{2Vp^2 dp d\Omega}{(2\pi\hbar)^3} . \tag{3.11.6}$$

Integrating over angle (in the isotropic case $\int d\Omega = 4\pi$), we have

$$d\mu = \frac{V\omega^2 d\omega}{\pi^2 c^3} = \frac{8\pi}{3} d\left(\frac{V}{\lambda^3}\right) . \tag{3.11.7}$$

Thus, in the isotropic case the occupation number n_{occ} is defined as

$$n_{occ} \equiv \frac{dN_\gamma}{d\mu} = \frac{\pi^2 c^3 \phi}{\omega^2} . \tag{3.11.8}$$

One can think of each electromagnetic-oscillation mode as a separate oscillator; then n_{occ} is the number of excited states of this oscillator. Also, since the energy of an oscillator can be changed only by a whole quantum, $\hbar\omega$, to the occupation number n_{occ} corresponds an energy $n_{occ}\hbar\omega$.

The number n_{occ} plays an important role in the physics of radiation. For example, induced (stimulated) emission in a laser differs from spontaneous emission by the factor $(n_{occ} + 1)$. In thermodynamic equilibrium radiation

$$n_{occ} = [\exp(\hbar\omega/kT) - 1]^{-1}$$

(the Planck formula). For a continuous spectrum, n_{occ} is statistically distributed over neighboring modes, though in reality we then mean the average value of n_{occ}.

During a slow expansion, one ought to expect that n_{occ} is conserved as an adiabatic invariant. And indeed, if one substitutes $\phi \sim n_{occ}\omega^2$ into the equation for ϕ, he obtains

$$\frac{\partial n_{occ}}{\partial t} - H\omega \frac{\partial n_{occ}}{\partial \omega} = \left.\frac{dn_{occ}}{dt}\right|_c = 0 . \tag{3.11.9}$$

Here d/dt_c means that the derivative is taken along a characteristic, i.e., subject to the condition

$$d\omega = -H\omega dt .$$

Thus, the macroscopic kinetic equation agrees with the quantum representation with respect to the number of photons per mode and the number of the quantum state.

Let us now return to the equation for ϕ. Adding the contribution due to sources, we obtain

$$\frac{\partial \phi}{\partial t} = -2H\phi + H\omega \frac{\partial \phi}{\partial \omega} + I(\omega, t) . \tag{3.11.10}$$

The equation for the spectral density $\xi(\omega, t)$ (see formula [3.11.2]) follows from this by an elementary calculation. The result is

$$\frac{\partial \xi}{\partial t} = -3H\xi + H\omega \frac{\partial \xi}{\partial \omega} + \Phi(\omega, t)$$
$$= -4H\xi + H \frac{\partial}{\partial \omega}(\omega\xi) + \Phi(\omega, t) . \tag{3.11.11}$$

Here $I(\omega, t)$ and $\Phi(\omega, t) = \hbar\omega I(\omega, t)$ are respectively the number of photons and the energy emitted by sources per unit volume per unit time in the frequency interval $d\omega$.

It is in essence clear that two methods are equivalent: that of the partial-differential equation (3.11.11) and that of summing the light, incident from the various layers of space surrounding any given point. The local approach, i.e., equation (3.11.11), is useful for the calculation of the general properties of the solution. For example, that the scattering of light (without absorption and without frequency change) in no way influences the spectrum is immediately apparent—an isotropic radiation field remains isotropic after scattering.

Integrating equation (3.11.11) with respect to ω, we obtain for the total energy density ϵ_γ the equation

$$\frac{d\epsilon_\gamma}{dt} = -4H\epsilon_\gamma(t) + Q(t) , \tag{3.11.12}$$

where

$$\epsilon_\gamma = \int_0^\infty \xi(\omega, t)d\omega ; \quad Q = \int_0^\infty \Phi(\omega, t)d\omega .$$

In formula (3.11.11) one sees that to the given solution of the equation with sources, one can always attach a solution of the homogeneous equation (that without sources). Such a solution is, in particular, the Planck distribution,

$$\xi_{\mathrm{Pl}}(\omega, t) = \text{const.} \frac{\omega^3 d\omega}{[\exp(\hbar\omega/kT) - 1]} , \tag{3.11.13}$$

with a temperature T depending on time according to

$$\frac{dT}{dt} = -HT . \tag{3.11.14}$$

Such an equilibrium Planck spectrum, with $T_{\mathrm{Pl}} = 2.7$ K, is observed at the present time as a consequence of the high temperature of matter at the early stages of expansion (see the following part of the book).

From equation (3.11.9) for n_{occ}, it is clear that one can obtain a rough estimate of the energy density due to single sources alone—without the contribution of the radiation of the hot model—by assuming in equation

(3.11.12) that the solution is quasi-stationary. Thus, neglecting the term in $\dot{\epsilon}_\gamma$ we obtain

$$-4H\epsilon_\gamma + Q = 0\,; \qquad \epsilon_\gamma = Q/4H\,. \tag{3.11.15}$$

Any function, appropriately compressed with respect to the frequency ω, is a solution of the source-free equation. That is, $n_{\text{occ}} = n_{\text{occ}}[\omega/\psi(t)]$, is a solution of (3.11.9) where

$$\psi \sim \exp\left(-\int H dt\right), \qquad \frac{d\psi}{dt} = -H\psi\,, \qquad \psi = (1+z)^{-1}\,.$$

For any function, i.e., for any form of the spectrum, the total number density of photons $\sim\int n_{\text{occ}}\omega^2 d\omega$ changes in proportion to $(1+z)^{-3}$, while the energy density changes in proportion to $(1+z)^{-4}$.

We proceed analogously for the equation giving the spectral density. The most rough approximation consists of eliminating the expressions $\partial\xi/\partial t$ and $\partial(\xi\omega)/\partial\omega$. The result is

$$\xi = \frac{\Phi}{4H}\,. \tag{3.11.16}$$

One can define the spectrum more precisely. Thus, neglecting the term $\partial\xi/\partial t$ but keeping the term with $\partial(\xi\omega)/\partial\omega$, we obtain

$$\xi(\omega, t) = \frac{\omega^3}{H}\int_\omega^\infty \frac{\Phi(\omega', t)}{\omega'^4}\, d\omega'\,. \tag{3.11.17}$$

In the stationary approximation it is also easy to construct a solution which takes account of the presence of dust in intergalactic space. This dust absorbs light, is heated, and thus emits its own thermal spectrum. Here, however, we will not discuss this.

The linear differential equation for the density of radiation, (3.11.12), can be solved exactly. One can write the solution in terms of a quadrature (definite integral) for any function $Q(t)$. The function $Q(t)$ appears linearly in the result, a fact that makes manifest the physical meaning of the problem. That is to say, the radiation density at a given instant t_0 is the sum of contributions $Q(\tau)d\tau$ of radiation from each small interval of time between τ and $\tau + d\tau$, where $\tau < t_0$. Each such contribution is appropriately transformed in the course of the expansion from the time τ to the time t_0. If one selects in this way the source contributions Q for the time from t_1 to t_0, then one must still add the contribution of the energy density which is already present at the time t_1. It is easy to verify that the expression

$$\epsilon_\gamma(t_0) = \int_{t_1}^{t_0} Q(\tau)\exp\left[-4\int_\tau^{t_0} H(t')dt'\right]d\tau + \epsilon_\gamma(t_1)\exp\left[-4\int_{t_1}^{t_0} H(t')dt'\right] \tag{3.11.18}$$

corresponds to these ideas and satisfies the equation exactly. For the sake of clarity, the variables of integration (τ, t') and the limits (t_0, t_1) are denoted by different letters.

The exponentials with the integrals have a very simple meaning. Since

$$H(t) = \frac{1}{a(t)} \frac{da(t)}{dt},$$

we have

$$\int_\tau^{t_0} H(t')dt' = \ln\left[\frac{a(t_0)}{a(\tau)}\right]; \qquad \exp\left[-4\int_\tau^{t_0} H(t')dt'\right] = \left[\frac{a(\tau)}{a(t_0)}\right]^4.$$

Since the energy density varies inversely with the fourth power of the scale factor, the energy density present or emitted at τ will fall by the factor $[a(\tau)/a(t_0)]^4$ by time t_0. This is in complete correspondence with the characteristics of thermodynamic radiation. The ratio of the scale factors is connected with the redshift,

$$a(\tau)/a(t_0) = 1 - \Delta(\tau) = [1 + z(\tau)]^{-1},$$

and so one may let Δ or z take the role of integration variable. At the same time, we employ expression (3.3.35) that connects the differential dt with $d\Delta$ or dz to obtain

$$\begin{aligned}
\epsilon_\gamma(t_0) &= \int_0^{\Delta_1} Q(\Delta)(1-\Delta)^4(1-\Delta)^{1/2}[1 + (\Omega_0 - 1)\Delta]^{-1/2}d\Delta \\
&\qquad + \epsilon_\gamma(t_1)(1-\Delta_1)^4, \\
\epsilon_\gamma(t_0) &= \int_0^{z_1} Q(z)(1+z)^{-4}(1+z)^{-2}(1 + \Omega_0 z)^{-1/2}dz \\
&\qquad + \epsilon_\gamma(t_1)(1+z_1)^{-4}.
\end{aligned} \tag{3.11.19}$$

Remember that Q is the instantaneous radiation power per unit physical volume. If one introduces instead the radiation power per unit mass, q^*, then it will be necessary to take account of the relation $Q = q^*\rho$:

$$\rho = \rho_0(1+z)^3 = \rho_0(1-\Delta)^{-3}; \qquad Q = q^*\rho_0(1+z)^3 = q^*\rho_0(1-\Delta)^{-3}.$$

One sometimes introduces Q_s, the radiation power per unit comoving volume. This quantity is defined so that at the present time, t_0, it coincides with Q. Evidently, $Q_s = Q(1+z)^{-3} = Q(1-\Delta)^3$. Correspondingly,

$$\begin{aligned}
\epsilon_\gamma(t_0) &= \int_0^{\Delta_1} Q_s(\Delta)(1-\Delta)(1-\Delta)^{1/2}[1 + (\Omega_0 - 1)\Delta]^{-1/2}d\Delta \\
&\qquad + \epsilon_\gamma(t_1)(1-\Delta_1)^4, \\
\epsilon_\gamma(t_0) &= \int_0^{z_1} Q_s(z)(1+z)^{-1}(1+z)^{-2}(1 + \Omega_0 z)^{-1/2}dz \\
&\qquad + \epsilon_\gamma(t_1)(1+z_1)^{-4}.
\end{aligned} \tag{3.11.20}$$

The assumption that Q, Q_*, and q^* depend directly on Δ or z, but not on t, and that they are expressed in terms of simple formulae as functions of Δ or z, is wholly natural since Δ and z are directly connected with the history. For example, under the simplest assumption, namely in the absence of evolution, $q^* = \text{const.}$, then

$$Q_* = \text{const.}, \qquad Q \sim (1+z)^3 \sim (1-\Delta)^{-3}.$$

In this case the integral converges even if one extends it to the horizon ($\Delta_1 = 1$, $z_1 = \infty$). This is just what is meant when one says that no Olbers paradox occurs in an expanding Universe (see, for example, Landau and Lifshitz 1973).

Let us consider another example. Suppose that the Universe is filled with ionized plasma. The emission satisfies $Q \sim n_e^2 T^{1/2}$, where T is the temperature and n_e the electron density. For constant temperature, $Q \sim (1+z)^6 \sim (1-\Delta)^{-6}$, and the integral would diverge as the horizon ($z_1 \to \infty$) is approached. If one includes the growth of the temperature, the divergence only becomes more pronounced. In the region of high temperatures, where $T \sim t^{-2/3}$, the dependence of dt on $d\Delta$ or dz changes in such a way that these conclusions remain valid.

In the hot model of the Universe, one assumes just this: that in early, prestellar stages, the matter exists as a completely ionized, homogeneously distributed plasma. Meanwhile, a definite radiation density is present because at high density there is thermodynamic equilibrium between the plasma and the radiation. This ensures that the radiation counterbalances the absorption of light, an effect which was not allowed for in the preceding discussion. Subsequently, the matter becomes neutral and ceases to interact with the radiation. The photon "gas" then expands without interacting with the matter. The contribution of the radiation from this stage of the expansion to the average density has a precise Planck spectrum corresponding to a temperature of 2.7 K. As was already noted, scattering does not play a role if, in the presence of scattering, the frequency and energy of the light do not change. More will be said about this in the next part of the book.

The partial differential equation (3.11.11) holds for the spectral density $f(\omega, t)$. One can solve this equation analogously to the method used for ϵ_γ by passing to Δ or z as integration variable (Doroshkevich and Novikov 1964). In the formal theory, the law for the change in frequency with redshift gives an equation with the characteristics of a partial differential equation. The solution was presented above in § 3.8, in connection with the analysis of the spectrum in the radio-frequency range.

In § 3.8 it was also shown that the optical depth of light due to scattering can conveniently be written in terms of dt/dz. In this section a discussion

of the method for the calculation of the average radiation density has been presented.

A review of the relevant observational data and also several concrete results of the calculations appear in § 5.2 and § 5.3.

3.12 Is the Redshift Unambiguously Explained by the Expansion of the Universe?

As history shows, the idea of a stationary Universe is quite alluring. However, this feeling is probably based on mental inertia. In the lifetime of civilization, or even in the lifetime of man, noticeable changes in the vast majority of cosmic systems have not occurred. Therefore, when observational proof of the redshift of spectral lines appeared, there followed many attempts to explain these shifts without employing the Doppler effect. Many authors would like to avoid schemes involving the Doppler effect and the mutual recession of galaxies. A picture encompassing an expanding Universe is so grandiose. And man is accustomed to low velocities. A much less unusual idea, one less likely to cause (or involve) a controversy, is that of an unevolving Universe. From this idea have followed numerous attempts to defend the idea of a stationary Universe, to give some sort of explanation to the "cosmological redshift," an explanation differing from the Doppler shift due to galaxy recession. Unfortunately, similar attempts are sometimes encountered even today.

In these explanations, use is made of the fact that the shift is indeed a redshift. If $\rho > \rho_{\text{crit}}$, the redshift will become a blueshift after $\sim 10^{10}$ yr. The aforementioned alternative explanations would go away automatically were this so. But now, in any case, $H > 0$ and the shift in the spectral lines corresponds to a decrease in the photon energy, i.e., to a loss by photons of a part of their energy along the path from distant objects to the Earthbound observer.

In this connection there still remains a question. How unambiguous is the interpretation of the redshift as due to the Doppler effect? Can another physical mechanism lead to a reddening of the photons? The first variant of an attempt at a different explanation is based on the GTR gravitational redshift. It is known that, in GTR, photons redden when they propagate from a region of strong gravitational potential to one with a weak potential. For example, photons moving upward near the surface of the Earth redden; conversely, photons moving downward become bluer. Indeed, this effect has been measured in the laboratory.

The effect of the reddening of photons in a strong gravitational field can in no way, however, explain the cosmological redshift. This is clear from the examination presented in § 3.5. First of all, given the present-day density, the effect is extraordinarily weak in a homogeneous Universe. Second, the shift is proportional to the square of the distance, and not to

of great power as is the case for the Hubble law. Third, and most important, the explanation involves the other sign. The shift must be blue—not red! Indeed, this is apparent from formula (3.5.10).

Of course, the gravitational contribution to the change in the frequency of photons is accounted for exactly in the formulae governing the cosmological redshift as the formulae of § 3.5 clearly demonstrate. This effect is of second order, however, and, in any case is not fundamental for the red shift. Frequently involved as an explanation of the redshift is the "tired light" idea: the idea that through some mechanism photons lose an amount of energy proportional to the distance they traverse. Two variants of this idea have arisen.

The first is that along the path from source to observer, the photons interact with the intergalactic matter and give to it a part of their energy. Such an explanation can be immediately refuted, for the interaction must have the character of scattering. The transfer of energy is thus accompanied by the transfer of momentum. At the same time, the direction of propagation of a photon must change. This in turn must lead to a blurring of the image of the source. But such a blurring is not observed in practice.

The second variant involves the spontaneous decay of photons. For example, a photon might emit a neutrino-antineutrino pair and thereby lose a small part of its energy. Now, in accordance with the laws of energy and momentum conservation, a photon can only emit particles with zero rest mass. Moreover, any such particles must move parallel to the photon's direction of motion.

Bronshtein (1934) showed, however, that if such a process occurs, it could long ago have been noticed in laboratory experiments. The essence of the argument is that the probability for spontaneous decay of photons must be inversely proportional to their frequency. This conclusion follows from the following considerations.

Recall, for example, the well-known relation between the energy and the lifetime for spontaneous decay for the μ meson. Suppose that τ_0 is the lifetime of a muon at rest. If the muon moves at the speed v, then its lifetime $\tau(v)$ is

$$\tau(v) = \tau_0[1 - v^2/c^2]^{-1/2} . \tag{3.12.1}$$

The energy of the moving muon is meanwhile

$$E = m_\mu c^2[1 - v^2/c^2]^{-1/2} , \tag{3.12.2}$$

where m_μ is the muon rest mass. The decay probability W is inversely proportional to the lifetime, $W \sim \tau^{-1}$. It then follows that

$$EW \sim E\tau^{-1} = \text{const} . \tag{3.12.3}$$

Since (3.12.3) does not explicitly contain the muon rest mass, it is a general

result which remains valid even in the case of the photon with zero rest mass.

For the question of photon decay, $E_\gamma = \hbar\omega$, so that

$$W_\gamma \sim E_\gamma^{-1} \sim \omega^{-1} . \tag{3.12.4}$$

This is the Bronshtein formula. In accordance with the principle of Lorentz invariance, this formula is universal, applying for muons as well as for photons. But if the decay probability of a photon were actually inversely proportional to the frequency, then radio-frequency photons would decay particularly rapidly. But nothing at all like this is observed. The redshift in the radio band is exactly the same as in the optical band. Indeed, this has been directly verified by observation of the 21 cm radio line in distant galaxies.

Further, the Bronshtein result would mean that there must be changes in Maxwell's equations for a static electric field. This is such a sensational prediction that the effect would certainly have been observed in the laboratory. But it has not been observed.

In spite of the considerations mentioned, a concrete variant of the tired-light hypothesis has been developed recently by Pecker, Roberts, and Vigier (1972). They suppose that photons lose energy by scattering from other photons. For intergalactic propagation it is assumed that scattering from RR plays a role. They cite authors who maintain that rays moving near the solar disk undergo a redshift that is intensified in accordance with the radiation density that is characterized by the temperature of the Sun.

These observations lie at the limit of accuracy and are therefore not proved. The weakness of the hypothesis is that it must be assumed that the photon-photon interaction conspires to change the energy but not the direction of the observed photon. Now, quantum electrodynamics does predict photon-photon scattering, but with negligibly small cross section at low frequencies (which include the optical and soft X-ray part of the spectrum). Without exception, the predictions of quantum electrodynamics are in accord with experiment. The explanation of Pecker *et al.* for the redshift is untenable.

Once again we emphasize that, even without a measurement of the redshift, the equations of mechanics guarantee that a homogeneous distribution of matter must be nonstationary (see § 1.2). The observations of the redshift in the spectral lines of galaxies then serve to support this result.

1 THE COSMOLOGICAL CONSTANT

11 Does the Cosmological Constant Differ from Zero?

The basis of the theory of gravity is the notion of the possibility of space-time curvature described in terms of a Riemannian metric. Remarkably, this idea leads almost uniquely to a definite form for the equations of GTR. One uses only the general and natural assumptions that, in the appropriate limits, the theory reduces to the special theory of relativity and Newton's theory of gravity.

All the same, as Einstein himself showed, one possibility for changing GTR is that of introducing into GTR a parameter which must be determined from observations or experiment. This parameter turns out to be important only on the scale of the Universe, and it therefore received the name "cosmological constant." It is denoted by Λ and has the dimension cm^{-2}.

In principle, one could determine Λ sufficiently accurately by laboratory methods. However, the dimension of Λ already makes it apparent that the effects related to this quantity are characterized by the dimensionless product Λl^2. Here l is a length characteristic of the experimental apparatus or the observation region. But extragalactic or cosmological observations performed by astronomers limit the value of this quantity to such a degree ($|\Lambda| \leq 10^{-55}\,\text{cm}^{-2}$) that laboratory determinations of it become hopeless. As was already noted in the Introduction, the question surrounding Λ, then, is an entirely new one that arises only upon passing to the gigantic scales of the Universe itself.

Before we write the relevant equations, let us dwell a bit on the history of the question and on several points of principle. Setting out to develop cosmology on the basis of GTR, Einstein considered it advisable to find a static solution with closed, three-dimensional spatial sections. First, it was assumed that this static character (time independence) corresponds to the great age of heavenly bodies, such as the Earth whose age of several billion years was already known. Second, the closed model was presumed to especially correspond to the physical ideas of Mach—Mach's principle (see Part V). That is, a closed model contains a finite amount of matter, and it is possible that this matter somehow selects a local inertial system of coordinates. Mach thought that the inertia of a body depends on the inter-

action of this body with the remaining matter. Such a viewpoint is more acceptable if the amount of remaining matter is finite.

All the calculations of the preceding chapters were carried out under the assumption $\Lambda = 0$, i.e., were carried out without considering the cosmological constant. A static solution does not appear among the solutions obtained, but the closed or open character of the geometry of a solution does indeed depend on the matter density.

In 1917 Einstein did not have nonstatic solutions available, for Friedmann only obtained them in the years 1922–1924. However, direct substitution of a static metric $g_{\mu\nu}$ for a closed model indicated that the equations could not be satisfied without modification. It turned out that a most natural change—introducing Λ—allows one to satisfy the equations. It still seemed that only certain very general principles are needed to specify the form of the equations and, what is even more important, to permit one to guess the structure of the Universe.

Then the work of Friedmann appeared. He indicated and examined all the variants: the open and closed models, both of which evolve, and also the static model, which is a particular solution when $\Lambda \neq 0$. After a short misunderstanding, Einstein recognized the theoretical correctness of this work.

Soon after 1929, by which time Friedmann had died, Hubble announced the discovery of the redshift law. This relation was immediately interpreted as confirmation of the evolving model of the Universe. Not only from a scientific point of view, but also from the point of view of the psychology of scientific creation (in the sense of "kicking a man who is already down"), it is highly instructive to see how the arguments in favor of the closed, static model lost their force simultaneous with Hubble's announcement. It also became clear that stars gradually expend their nuclear energy: although they can shine for a long time (the Sun for $\sim 10^{10}$ years, a massive star for $\sim 10^6$ years), they cannot shine forever. From this viewpoint as well, the static model is not suitable. In addition, it was later shown that the static model is not stable with respect to small density perturbations.

True, it once seemed that models without the Λ term give an excessively short time scale—the time from the singularity to the present—of several billion years in all (i.e., Hubble's constant seemed to have a value of 500 km s^{-1} Mpc^{-1}). This time is even less than the age of the Earth! Thus, the introduction of the Λ term was thought to be advisable to stretch out the expansion time of the Universe. In the following years, however, astronomers realized that the distances to the galaxies had been seriously underestimated. This revision led to the value $H_0 \approx 75$ km s^{-1} Mpc^{-1}. The contradictions with the short time scale thus vanished, as did the need for a Λ term.

Mach's principle gradually faded from view and cannot serve as an

opponent in favor of the closed Universe. Nevertheless, the critical analysis conducted by Mach, based on classical physics, was psychologically useful as a preparation for new theories. Note, however, that electrodynamics, SR, and GTR as well, all of which are constructed on the foundations (not on the ruins!) of classical physics, are local theories. They do not encompass any action at a distance, something implicit in Mach's principle.

After Friedmann's discovery of the nonstationary solution, Einstein announced that the introduction of Λ was an unnecessary complication of the theory and that no data existed which would demand the introduction of Λ. Landau and Lifshitz express the same point of view in their *Classical Theory of Fields* (1962; see page 375).

One cannot deny, however, that such an approach is subjective. Although astronomical data did not demand that Λ differ from zero, they also did not deny the possibility that $\Lambda \neq 0$. In any case Λ must be sufficiently small in order that the theory not be in conflict with the data about "our neighborhood" (up to 1000–1500 Mpc in distance or $3\text{–}5 \times 10^9$ years in time scale) yet be sufficiently large to radically influence the large-scale structure and evolution of the Universe.

Given the tendency toward objectivity and completeness, this already could have been reason enough for the theory with a nonzero Λ to be located (at least in small type—in an appendix). But sensible, calm decisions are often made only under the pressure of extraordinary circumstances, in a situation of fire and/or flood and panic.

Quasars played the role of the torch. In 1967, data appeared—their statistical significance is, to this day, not clear, and most likely comes to naught—about a concentration of quasars at the specific redshift value $z = 1.95$ (see Burbidge and O'Dell 1973). An explanation of this result was advanced within the realm of a certain class of cosmological models. Such a model is, for a long time, almost "static" at a specified value of z because of the effect of a nonzero cosmological constant. Stormy discussions followed, regarding the arguments for and against such an explanation. In this same connection, questions arose about the relation of GTR (in particular of the cosmological constant) and the theory of elementary particles. This group of questions is considered in Volume I (see also ch. 24 of this volume), so we will not return to them here.

Knowledge of the basic properties of the theory with $\Lambda \neq 0$ became necessary in order to understand the relevant articles published in the journals. We therefore consider it advisable to present here the basic formulae concerning homogeneous, isotropic models with a nonzero Λ term. In addition, a reading of the subsequent material might well help in the formulation of theoretical observational and research programs that will explain why $\Lambda = 0$. In all of the remaining parts of the book, however, we take $\Lambda \equiv 0$.

4.2 Cosmological Models with the Λ Term

GTR permits the introduction, in covariant form, of the so called cosmological constant Λ. The equations then have the form

$$R_{ik} - \tfrac{1}{2}g_{ik}R = \frac{\kappa}{c^2}T_{ik} + g_{ik}\Lambda\,, \tag{4.2.1}$$

where R_{ik} and R characterize the curvature of spacetime (see § 2.1), T_{ik} is the stress-energy tensor of the matter, $\kappa = 8\pi G/c^2$ is the Einstein gravitational constant, and $g_{ik}\Lambda$ is the cosmological term.

From these equations we see that the new term is equivalent to an additional term in the stress-energy tensor. This term contributes an energy density

$$\epsilon_\Lambda = \frac{c^4\Lambda}{8\pi G}\,, \tag{4.2.2}$$

and a pressure

$$p_\Lambda = -\epsilon_\Lambda = -\frac{c^4\Lambda}{8\pi G}\,. \tag{4.2.3}$$

One can introduce, as well, the equivalent mass density

$$\rho_\Lambda \equiv \frac{\epsilon_\Lambda}{c^2} = \frac{c^2\Lambda}{8\pi G}\,. \tag{4.2.4}$$

The quantities introduced do not depend on the density of particles. One can therefore talk about ρ_Λ, ϵ_Λ, and p_Λ as though they represent the mass density, energy density, and pressure of empty space—of the vacuum. We will analyze their effects on the behavior of the Universe in the Newtonian manner as was done in chapter 1. Such an analysis will also permit us to explain why Einstein named Λ the cosmological constant.

At the same time, we will show that the cosmological constant term contributes $\epsilon_\Lambda + 3p_\Lambda = \epsilon_\Lambda - 3\epsilon_\Lambda = -2\epsilon_\Lambda = -c^4\Lambda/(4\pi G)$ to the equations for the acceleration (see eqs. [1.6.1] and [2.1.8]). Further, we will write equation (1.6.4) for the "kinetic energy" with the same value for the constant as follows from GTR (see eq. [2.1.9]) when the scale factor is chosen as the radius of the model. Thus,

$$\frac{1}{2}\left(\frac{da}{dt}\right)^2 = \frac{4\pi G}{3c^2}a(\epsilon + \epsilon_\Lambda) - \frac{Kc^2}{2}\,. \tag{4.2.5}$$

Here, $K = -1$ for the hyperbolic metric (formula [2.3.2]) of an open model; $K = +1$ for the spherical metric (formula [2.2.6]) of a closed model; and $K = 0$ for the flat, Euclidean metric (formula [2.5.2]), with the specification that one write b in place of a. In the equation above, ϵ and p without subscripts refer to ordinary matter—stars, intergalactic gas, and radiation. Obviously, $\epsilon > 0$ and $p > 0$. Notice that ϵ_Λ and p_Λ satisfy an energy conservation law (first law of thermodynamics) in the form $dE_\Lambda = d(\epsilon_\Lambda V) = p_\Lambda dV$ with $p_\Lambda = -\epsilon_\Lambda = \text{const}$. This agrees completely with

the interpretation of these quantities as the energy density and pressure of the vacuum. Moreover, as is noted in Volume I, the values of ϵ_Λ and p_Λ are Lorentz invariant: the same in all coordinate systems moving with respect to one another.

Instead of equations (2.1.8) and (2.1.9), we now have

$$\begin{aligned}\frac{d^2a}{dt^2} &= -\frac{4\pi G}{3c^2}\,a(\epsilon + 3p + \epsilon_\Lambda + 3p_\Lambda) \\ &= -\frac{4\pi G}{3c^2}\,a(\epsilon + 3p - 2\epsilon_\Lambda)\,,\end{aligned} \tag{4.2.6}$$

$$\frac{1}{2}\left(\frac{da}{dt}\right)^2 = \frac{4\pi G}{3c^2}\,a^2(\epsilon + \epsilon_\Lambda) - \frac{Kc^2}{2}\,. \tag{4.2.7}$$

The most important property of the new equations is that they allow for a static solution. If

$$\epsilon_\Lambda = \frac{\epsilon + 3p}{2}\,, \qquad K = 1 = \frac{4\pi G}{c^4}\,a^2(\epsilon + p)\,, \tag{4.2.8}$$

then

$$\frac{d^2a}{dt^2} = 0\,, \qquad \frac{da}{dt} = 0\,, \qquad a = \text{const}\,. \tag{4.2.9}$$

follow from equations (4.2.6) and (4.2.7). Notice that the static solution cannot exist without the cosmological term since $\epsilon > 0$ and $p > 0$ for matter. Equation (4.2.6) is thus not compatible with $d^2a/dt^2 = 0$ if the Λ terms are absent. When $\Lambda \neq 0$, the static solution appears by virtue of the negative pressure p_Λ which accompanies and exceeds the positive energy density ϵ_Λ in the expression $\epsilon_\Lambda + 3p_\Lambda = \epsilon_\Lambda - 3\epsilon_\Lambda = -2\epsilon_\Lambda$. It is further evident that the static solution must be closed. The condition $d^2a/dt^2 = 0$ gives $\epsilon + 3p - 2\epsilon_\Lambda = 0$. From this it follows that $\epsilon + \epsilon_\Lambda$, the quantity on the right-hand side of the equation for $(da/dt)^2$, is positive definite, namely, $\epsilon + \epsilon_\Lambda = \epsilon + (\epsilon + 3p)/2 = 3(\epsilon + p)/2 > 0$. Consequently, if $(da/dt)^2 = 0$, then $K = +1$ and the model is closed. Its radius is expressed in terms of the density and pressure of matter. Such a solution, called the Einstein cosmological model, realized Einstein's wish for a closed and static Universe. This interpretation closely follows that of Bondi (1961).

Let us turn to the cosmological model with $\Lambda \neq 0$ that was advanced in connection with research on quasars. At first Petrosian, Salpeter, and Szekeres (1967) introduced a nonzero Λ in order to explain the redshift–apparent-magnitude relation for quasars. Soon after this Shklovsky (1967) and Kardashev (1967) proposed a cosmological model with the Λ term chosen such that the expansion has a long delay at a specific radius. Indeed, there is some indication of a concentration of quasars at the redshift $z_{\text{plat}} = 1.95$. A possible explanation is that this redshift is achieved at a radius a_{plat} for which equation (4.2.8) is approximately satisfied. At this radius,

the Universe for a long time hardly expanded at all. The connection between the radius and the redshift, $a_0/a_{\rm plat} = 1 + z_{\rm plat}$, is preserved even when $\Lambda \neq 0$. Thus, these authors set forth a model which, over a long time period, behaves very much like the static Einstein model. The model is consequently closed. The "plateau" condition for the function $a(t)$ at the specific redshift $z_{\rm plat} = 1.95$, along with the present-day value of Hubble's constant H_0, completely determines the basic properties of the model. Below we suppose that one can neglect the pressure due to any matter or radiation present.

For this model it is easy to show that

$$\Omega_0 = \frac{2}{z^2_{\rm plat}(3 + z_{\rm plat})} \approx 0.1 , \quad \frac{\rho_\Lambda}{\rho_c} = \frac{(1 + z_{\rm plat})^3}{z^2_{\rm plat}(3 + z_{\rm plat})} \approx 1.4 . \qquad (4.2.10)$$

Recall that $\rho_c \approx 2 \times 10^{-29} h_0^2$ g cm^{-3} where $H_0 = 100 h_0$ km s^{-1} Mpc^{-1}. Then we obtain

$$\begin{aligned} \rho_\Lambda &\approx 2.7 \times 10^{-29} h_0^2 \text{ g cm}^{-3} , \\ \epsilon_\Lambda &\approx 2.5 \times 10^{-8} h_0^2 \text{ ergs cm}^{-3} , \\ \Lambda &\approx 5.0 \times 10^{-56} h_0^2 \text{ cm}^{-2} . \end{aligned} \qquad (4.2.11)$$

The present value of the matter density would be $\rho_0 \approx 2 \times 10^{-30} h_0^2$, and the present value of the scale factor would be

$$a_0 = \left(\frac{c}{H_0}\right) \frac{(1 + z_{\rm plat})}{z_{\rm plat}} \left(\frac{3}{3 + z_{\rm plat}}\right)^{1/2} \approx 1.2 \times 10^{+28} h_0^{-1} \text{ cm} . \qquad (4.2.12)$$

At the plateau ($z = z_{\rm plat} \approx 1.95$) we have

$$\begin{aligned} \rho_{\rm plat} &= (1 + z_{\rm plat})^3 \rho_0 \approx 5.2 \times 10^{-29} h_0^2 \text{ g cm}^{-3} , \\ a_{\rm plat} &= (1 + z_{\rm plat})^{-1} a_0 \approx 4 \times 10^{27} h_0^{-1} \text{ cm} . \end{aligned} \qquad (4.2.13)$$

One can take any value for the "coasting period" on the plateau. This time is proportional either to $\ln [\rho_0/(\rho_0 - \rho''_0)]$ or $\ln [\Lambda/(\Lambda - \Lambda'')]$, where the quantities with two primes are, in essence, the values of the parameters which would yield the strictly static model. Kardashev takes the duration to be of order 7×10^{10} yr, so that $(\rho_0 - \rho''_0)/\rho_0$ is necessarily of order 10^{-4}.

Sunyaev (1970) has called attention to the difficulties with models having a long plateau period. A closed model with a long delay in the expansion possesses a distinctive peculiarity: during the coasting period, a light or radio signal has sufficient time to circumnavigate the Universe several times. For example, after a time equal to $\pi a_{\rm plat}/c \approx 4 \times 10^{17}$ s, a photon emitted from the anticenter arrives at the origin. The anticenter is the point where $r = \pi$ in the metric $ds^2 = c^2 dt^2 - a^2(t)[dr^2 + \sin^2 r\,(d\theta^2 + \sin^2 \theta d\phi^2)]$. After twice as much time, the photon will circumnavigate the whole Universe and return to the point of emission. Formally, focusing will

or at $r = 0$ or $r = \pi$. For a source at $r = \pi$, therefore, the angular and bolometric distances approach zero. The energy flux received then strongly depends on whether, by chance, a source of radiation is located at the anti-center. This also applies to a source that, in the distant past, is at the same place where our Galaxy is now located.

Since observations do not support this focusing effect, many authors have discussed the reasons for the evident destruction of the exact focusing characteristic of a strictly homogeneous and isotropic model. When all is said and done, though, the concept of a cosmological model with $\Lambda \neq 0$ and thus with a plateau in the function $a(t)$ has, to a significant degree, lost its attractiveness at present.

Now we shall examine the general formulation of the problem of the dynamics and geometry of a model with $\Lambda \neq 0$ in order to classify all solutions. For convenience in making comparisons with observations we shall assume that the present-day values of the Hubble constant and the matter density are known. We shall also neglect the pressure of matter. Further, suppose that either ρ_Λ or the value of the acceleration related to ρ_Λ by equation (4.2.6) is also known.

The character of the solutions depends only on dimensionless quantities. We again introduce the dimensionless density Ω_0 and the dimensionless (but the square of the Hubble constant is included) deceleration parameter $q_0 = H_0^{-2} a_0^{-1} (d^2a/dt^2)_0$. As long as theories with $\Lambda = 0$ and $p = 0$ are examined, the quantities q_0 and Ω_0 are equivalent: $q_0 = \Omega_0/2$. However, when $\Lambda \neq 0$, we easily find that $q_0 = \Omega_0/2 - 8\pi G\rho_\Lambda/(3H_0^2)$, so that the difference $\Omega_0/2 - q_0 = 8\pi G\rho_\Lambda/(3H_0^2)$ characterizes Λ. The quantity $(\Omega_0/2 - q_0)$ is the dimensionless density corresponding to the cosmological constant:

$$\lambda \equiv \frac{\rho_\Lambda}{\rho_c} = \frac{\Lambda c^2}{3H_0^2}. \tag{4.2.14}$$

Let us further introduce $x \equiv (1 + z)^{-1} = a(t)/a_0$, $\tau = tH_0$, and $\alpha^2 = |a_0^2 H_0^2/c^2|$, the last being the dimensionless scale factor. When we consider the present time, $\alpha^2 = \pm a_0^2 H_0^2/c^2$ appears in the equations; the positive sign relates to the closed model, the negative sign to the open or hyperbolic model. We transform equation (4.2.5) to the form

$$\frac{1}{2}\left(\frac{dx}{d\tau}\right)^2 = \frac{x^2}{2}\left(\frac{\Omega_0}{x^3} + \lambda\right) - \frac{1}{2\alpha^2}. \tag{4.2.15}$$

The equation obtained has the form of an energy conservation law for the motion of a material point, with time-dependent coordinate $x(\tau)$, in a potential depending upon x. One can further transform this equation by introducing q_0 in place of λ and by taking account of the characterization of the present-day expansion velocity by H_0. When $x = 1$, $[da/dt]_{a=a_0} = H_0 a_0$, from which it follows that $dx/d\tau = 1$ for $x = 1$. Owing to this, the

constant α^{-2} is not arbitrary, but is expressed in terms of Ω_0 and λ or in terms of Ω_0 and q_0. Thus, we finally obtain

$$\left(\frac{dx}{d\tau}\right)^2 = \frac{\Omega_0}{x} + \left(\frac{\Omega_0}{2} - q_0\right)x^2 + 1 + q_0 - \tfrac{3}{2}\Omega_0 . \qquad (4.2.16)$$

In this form the equation applies to all three geometric variants: the closed, hyperbolic, and flat models. The solution describing our Universe passes through the point $x = 1$. To obtain all the detailed information about the dependences of the brightness and the number of galaxies on redshift, one would first have to integrate equation (4.2.16), and then integrate the equation for photon propagation: $ds = 0$ or $cdt = a(t)d\tau$. The relations obtained are rather complicated, especially when the light executes repeated circumnavigations of the Universe. Many articles describe such work; see, for example, that of Refsdal and Stabell (1966) and Petrosian (1974). Here, however, we shall be content with studying the general character of the evolution.

For this purpose, construction of the curves of $f(x)$ is sufficient: $f(x)$ is the right-hand side of the expression (4.2.16). If $f(x) > 0$ in the interval $0 < x < 1$, then the motion from $x = 0$, $\tau = 0$ to $x = 1$, $\tau = \tau_1$ occurs monotonically and there is, in the past, a singularity. This evolution is qualitatively similar to the evolution of a model with $\Lambda = 0$ for which $f(x) = \Omega_0/x + 1 - \Omega_0$ and for which $f(x)$ satisfies the condition $f(x) > 0$ for $x < 1$. Let us call this evolution variant I.

Another case is possible, namely, when in the region $x_1 < x < x_2$, $f(x) < 0$. This means that the indicated region is forbidden, that there exists an energy barrier. In this case the scale factor is infinite in the past. The Universe first contracts from $x = \infty$ to a definite x_2, then expands from x_2 (reflection of the particle from the energy barrier), and at the present time continues in an expansion phase. Such a variant (call it II) is hardly suitable for a description of our Universe. No singularity occurs in it, nor does a period of high matter density. Recall that it is during a period of high density that the equilibrium spectrum of the so-called relic (or primordial) radiation arises. As is shown later, the RR spectrum is inexplicable within variant II. In order for a delay at a period of high density to occur, a high value for Λ, incompatible with observations, would be necessary.

One can analogously examine the future of the Universe. It depends on the behavior of $f(x)$ for $x > 1$. If $f(x) > 0$ for $x > 1$, then unbounded expansion takes place (variant III). If $f(x) = 0$ for a certain $x_3 > 1$, then the present state, $x = 1$, and $dx/d\tau = 1$ will continue to expand until $x = x_3$, at which time a change to contraction occurs. This contraction then continues to a singularity—to $x = 0$ (variant IV).

All of the information about the different variants of geometry and evolution for $\Lambda \neq 0$ is assembled in Figure 24. The coordinates are quantities that are in principle but, alas, not in practice—at least today—obtainable

mean of observations of immediately surrounding regions. They are the deceleration parameter q_0 along the abscissa and the dimensionless mean density Ω_0 along the ordinate. Remember that we consider the pressure negligible, q_0 can be positive or negative, and $\Omega_0 > 0$.

The line OAA' corresponds to $\Lambda = 0$—the case examined in detail above. The dashed lines parallel to it correspond to $\Lambda = \text{const}$. Those labeled lines with $\Lambda > 0$ are to the right of OAA', and those with $\Lambda < 0$ are to the left of OAA'. The line BAF corresponds to the spatially flat model ($k = 0$) and satisfies

$$\Omega_0 = \tfrac{2}{3}(1 + q_0) . \qquad (4.2.17)$$

The line divides the solutions corresponding to a closed model—to the left of BAF—from those for an open model—to the right of BAF.

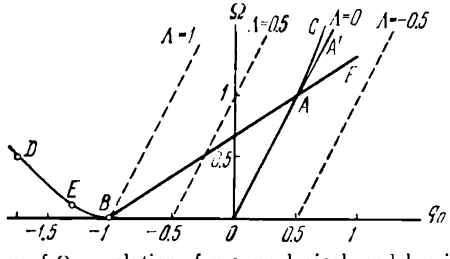


Fig. 24. Ω–Λ diagram of Ω–q_0 relations for cosmological models with various values of the cosmological constant Λ.

Now turn to the different types of evolution. The line BED separates two types of past: variant I to the right of BED and variant II to the left. The model of Kardashev, with a long delay, is represented by a point very close to point E but a little higher and more to the right. Finally, the line OAC separates two types of future. The equations predict unbounded expansion, variant III, for points to the left of OAC, while those to the right of OAC predict the behavior of variant IV. In the models with no Λ term, the future and the geometrical type are uniquely connected. The closed model evolves in accordance with variant IV, the flat or open models in accordance with variant III. As is apparent from Figure 24, this connection no longer applies when $\Lambda \neq 0$. For example, the model is geometrically (three-dimensionally) closed in the region above $GBAC$; nevertheless, unbounded expansion is in prospect. On the other hand, for points in the region $HOAF$, which represent a hyperbolic model, a transition from expansion to contraction is in prospect.

In conclusion, notice a curious, particular solution corresponding to the point B. This represents a flat model without matter and with the metric

$$ds^2 = c^2dt^2 - e^{2Ht}[dx^2 + dy^2 + dz^2] ;$$

with $\lambda = 1$ and $\Lambda = 3H_0^2/c^2 = 3 \times 10^{-56} h_0^2$ cm $^{-2}$ as necessary conditions. This metric coincides with that for the steady-state Universe. In this variant the expansion occurs without bound over time, both in the past and in the future. The Universe is not static but is stationary in the sense that all its local properties are independent of time. Formally adding to t a constant and changing the scale of the dimensionless coordinates return us to the original situation. Though literally not applicable to a Universe with matter (when $\Omega \neq 0$, $\rho_0 \neq 0$), this solution nevertheless has inspired several attempts at a description of an eternal, stationary Universe. Notwithstanding a negative attitude toward these attempts, we will examine them later.

Appendix to Part I

The Semiclosed Model and the Physical Interpretation of the Kruskal Metric

The closed model with $K = 1$ possesses a group of remarkable properties. For example, its total energy or, equivalently, its mass, identically equals zero; also, its total electric charge also equals zero (see Landau and Lifshitz 1962). For an understanding of these properties it is useful to examine the case of a homogeneous distribution of matter that ends at some distance from the origin of coordinates; farther from the origin there is only vacuum. Of course, this is simply the case of an isolated body. To appreciate the content of this appendix the reader must have some familiarity with the theory of the spherically symmetric gravitational field, the Schwarzschild metric, and the Kruskal-coordinate description of this metric. One can read about these subjects in Volume 1 of this book, *Stars and Relativity*.

For the closed model there is an instant of greatest expansion when the velocity is instantaneously equal to zero and the matter is instantaneously at rest. We will be examining this instant t_m and, for simplicity, shall limit ourselves to the case where the matter consists of pressure-free dust, so that $p = 0$. Remember (see § 1.1) that the evolution of a sphere of arbitrary radius, and all its properties, in no way depend on the presence or absence of surrounding matter. Such matter has no gravitational influence on the matter inside the sphere. Thus, we will examine a sphere that represents part of a closed Universe and will suppose that all the exterior matter, the matter outside of the radius R, is absent. We are thus discussing a sphere of dust in empty space.

Far from the sphere the gravitational field is weak, and at infinity space is Euclidean. Near the sphere the deformation of space depends on the strength of the gravitational field. For a fixed density, the strength of the field depends on the size of the sphere. As long as the radius, R, of the sphere is small, the curvature of the space in it is not significant since $a = [3c^2/(8\pi G\rho)]^{1/2} \gg R$ (see Fig. 25). But if the radius of the sphere increases, then the effect of the curvature on the structure of the sphere grows. As we saw in § 2.2, the surface area of the sphere is $S = 4\pi a^2 \sin^2 r$ and its volume is

$$V = 4\pi a^3[\tfrac{1}{2}r - \tfrac{1}{4}\sin 2r] ,$$

where $R = ar$. Finally, after an increase of the radius to a value greater than $R = a\pi/2$, there appears in the geometry of the interior of the sphere a peculiarity that is characteristic of the closed model. The area of the sphere decreases as its radius grows if it lies beyond the maximal "equatorial sphere." What are the properties of the space outside such a sphere, a "semiclosed world" (O. Klein 1961; Zel'dovich 1962*a*; Novikov 1962, 1963,

1964*a*; Harrison, Thorne, Wakano, and Wheeler 1965)? What does a semi closed world look like to an external observer?

The external gravitational field of a sphere must be spherically symmetric It is known (see, for example, Volume 1) that the only spherically symmetric gravitational field is that described by the Schwarzschild metric. Moreover, this field depends only on one constant m, the exterior gravitational mass, or total energy, of the sphere. The Kruskal metric describes the complete

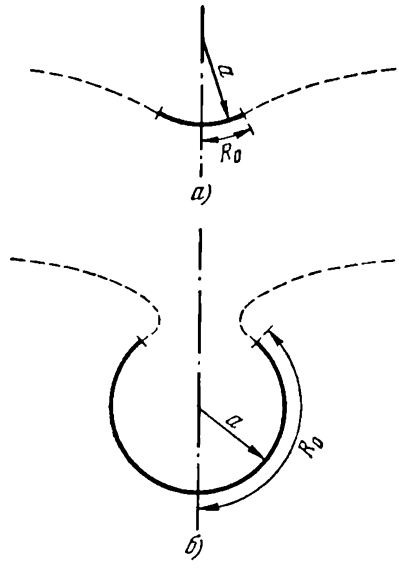


FIG. 25.—A representation of the curvature of space in the neighborhood of an isolated spherical body. Rotation of the curves shown around the vertical line gives surfaces that constitute a two-dimensional analog of the curved space near such a body. More specifically, these surfaces have the same geometry as an equatorial slice ($\theta = \pi/2$) through the curved space of the body. Drawing *a* is for the case where the radius of the body is much less than the spatial radius of curvature which the body's density generates: $R_0 \ll a = [3c^2/(8\pi G\rho)]^{1/2}$. Drawing *b* describes a "semiclosed world" with $R_0 > (\pi/2)a$.

structure of the Schwarzschild field (see § 3.14 of Volume 1), and the formulae for this metric appear at the end of this appendix. Here we will consider the question posed above qualitatively, making use of deductions from the formulae.

We will examine a sphere at the instant of greatest expansion and will consider changes in its radius, and consequently its mass, at this instant. The larger the mass, the larger the gravitational radius of the body. At the end of this appendix it is shown that this gravitational radius is $r_g^* = a_{\max} \sin^3 r$. At the time of maximal expansion of the surface of the sphere, the radius, defined as $[S_{\max}/4\pi]^{1/2}$, is evidently $r_{\max}^* = a_{\max} \sin r$. Consequently, the ratio, $r_g^*/r_{\max}^* = \sin^2 r$, increases with increasing size of the

sphere. As long as $r < \pi/2$ and $r_{max}{}^* > r_g{}^*$, a semiclosed model does not emerge. For such a sphere, the spacetime is shown in Figure 26 for the entire duration. Along the vertical axis of this diagram is the proper time of particles, and along the horizontal axis their Lagrangian coordinate R. The hatched region is occupied by matter. In vacuum, this system is continuously modeled by means of free test particles (for details, see § 3.14 of Vol. 1). The lines $r^* = r_g{}^*$ are the world lines of the gravitational radius. Ordinary spacetime far from the sphere, (Euclidean at infinity), is located to the right in the region R_1. The evolution of a sphere proceeds in the following manner as viewed by an external observer. The sphere expands from a state

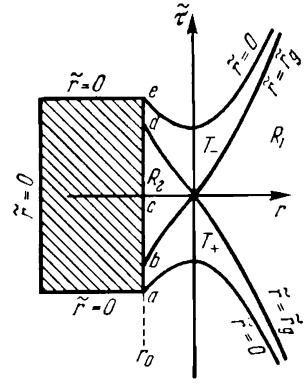


Fig. 26. The spacetime of a semiclosed world. The shaded region represents a matter-filled sphere.

of infinite density with $R = 0$ (point a). It then emerges from below its gravitational radius (at point b) and achieves the maximal size $R_{max} = r_{max}{}^*$ (at point c). Finally, it contracts, reaching the gravitational radius after a finite proper time (point d), and later $R = 0$ (at point e). To an external observer, the approach to point d extends over an infinite time (see § 3.14 of Vol. 1).

If one takes the size of the sphere such that $r = \pi/2$, then its energy is just sufficient for it to expand to $r_{max}{}^* = r_g$ and then contract (see Fig. 26).

If $r > \pi/2$, the surface of the sphere (now a semiclosed Universe), in the view of an external observer, never emerges from the Schwarzschild sphere. The surface of the sphere emerges from the T_+ region, from the Schwarzschild sphere, but not as seen by an external observer at the right of the R_1 region. It emerges only in a region to the left of the "throat" (see Fig. 26)—at the left of the R_2 region. The surface then achieves a maximum size and subsequently contracts. In the vacuum outside the semiclosed Universe the Kruskal metric with a throat describes the space. The radius of the throat changes with time; it is at its "widest" at the instant of greatest expansion

(see § 3.14 of Vol. 1). An external observer can receive information from the semiclosed Universe only about the early stages of expansion when the surface of the sphere is still located in the T_+ region (see Fig. 26).

Do semiclosed models have any relation to reality? We at present are unaware of such objects (see the analysis in § 14.5 of Vol. 1 and Markov 1973). But independent of the reality of their existence, an examination of the properties of semiclosed models has much methodological interest. Especially interesting are the external gravitational manifestations of a semiclosed Universe.

Thus, let us examine the properties of larger and larger (with respect to linear dimension) spheres, all of which have the same density, at the instant of maximal expansion. What is the mass of such spheres? Although static only for an instant, the gravitational mass m of a static configuration is less than the sum of the masses M of its constituent particles. This is due to the gravitational mass defect. But it is the gravitational mass m that can be measured by a distant observer through the gravitational field.

The gravitational mass can be expressed in the following form:

$$m = \frac{c^2 r_0^*}{2G} = \frac{c^2 a_{\max} \sin^3 r}{2G} = \frac{4\pi}{3} \rho a_{\max}^3 \sin^3 r \,. \qquad (4.A.1)$$

Meanwhile, the sum of the rest masses, M, is, by definition, given by the formula

$$M = \rho V = 4\pi\rho a^3 [\tfrac{1}{2} r - \tfrac{1}{4} \sin 2r] \,. \qquad (4.A.2)$$

For a small sphere with $r \ll 1$, both masses naturally coincide since the gravitational defect is small. Thus,

$$M \approx m \approx \frac{4\pi}{3} a_{\max}^3 r^3 \rho \,.$$

With increasing r, the sum of the masses of the constituent particles, M, grows as the volume V. The gravitational mass m grows more slowly because of the mass defect. The quantity m achieves a maximum for $r = \pi/2$:

$$m_{\max} = \frac{4\pi}{3} a_{\max}^3 \rho = \frac{2M}{3\pi} \,.$$

With further growth in r the mass m decreases. The addition of a new layer of matter produces an additional mass defect greater than the proper mass of this layer. For $r \to \pi$, $m \to 0$. Thus, the mass tends to zero as the configuration tends to closure. In short, one can say that the mass of a closed model equals zero. In such a model the gravitational mass defect equals the proper mass itself: $M = \rho V_{\text{tot}} = 2\pi^2 a_{\max}^3 \rho$. For another approach to the

distribution of the mass of a closed model that gives the same result, $m = 0$, Landau and Lifshitz (1962), § 107.

In a closed Universe the total electric charge also vanishes. Nonzero total electric charge would be possible only if Maxwell's equation were not correct. Why? Charge is the source of the electric field. In Euclidean space the lines of force of the electric field emanate from a charge and either gather together at a charge of the opposite sign or extend to infinity. But in a closed Universe there is no infinity. Thus, if we specify an arbitrary distribution of electric field $\boldsymbol{E}$ in a closed model, we find the corresponding charge density ρ_e by means of the equation

$$\nabla \cdot \boldsymbol{E} = 4\pi\rho_e .$$

In all cases the total electric charge is zero,

$$Q \equiv \int \rho_e dV = 0 ;$$

for if there is no infinity, then the lines of force beginning at one charge necessarily end at another charge of the opposite sign, neutralizing the first.

For a more detailed analysis of this and other properties of the closed models, see the study of Markov (1973).

In conclusion, we present the formulae which describe the properties of the closed Universe (see Novikov 1964) making use of the solution of Tolman (see § 3.13 of Vol. 1). The interval takes the form

$$ds^2 = a^2\{d\tau^2 - e^{\lambda}dr^2 - [r^*(\tau, r)]^2(d\theta^2 + \sin^2\theta d\phi^2)\} . \qquad (4.A.3)$$

Here, a is a constant with the dimensions of length that sets the scale, and all of the quantities inside the brackets are dimensionless. Tolman's solution is then

$$r^* = \Phi(r) = [f(r)]^{-1}[f(r)(r^*)^2 + F(r)r^*]^{1/2} + [-f(r)]^{-3/2}F(r)\sin^{-1}\left[\frac{-r^*f(r)}{F(r)}\right]^{1/2} , \qquad (4.A.4)$$

$$e^{\lambda} = [1 + f(r)]^{-1}(dr^*/dr)^2 , \qquad a^2\frac{8\pi G}{c^2}\rho = \frac{dF(r)/dr}{(r^*)^2(dr^*/dr)} .$$

The functions $f(r)$, $F(r)$, and $\Phi(r)$ are arbitrary and are determined by the conditions of a given problem. Suppose, then, that the region filled with matter, the region $r < r_0$, represents a part of the closed model. Then, passing from the expression (2.1.10) for the interval, we obtain

$$f(r) = \sin^2 r , \qquad F(r) = \sin^3 r ,$$
$$\Phi = \pi/2 \quad \text{for} \quad 0 \le r \le r_0 ; \qquad (4.A.5)$$
$$a = a(t)_{\max} , \quad \text{where} \quad a(t)_{\max} = \left[\frac{3c^2}{8\pi G\rho_{\min}}\right]^{1/2} .$$

If $\pi/2 < r_0 < \pi$, the matter forms a semiclosed Universe. Outside the sphere (for $r > r_0$) is vacuum. In order to match this solution (4.A.5) with the Kruskal metric (see § 3.14 of Vol. 1) for the vacuum, one must set

$$f = -[(r + c_1)^2 + c_2]^{-1} \sin^6 r_0 , \qquad F = \sin^3 r_0$$
$$\Phi = \frac{\pi}{2}\left[\frac{(r + c_1)^2 + c_2}{\sin^4 r_0}\right]^{3/2} , \tag{4.A.6}$$

for $r > r_0$. Here $c_2 = \sin^6 r_0$ and c_1 is expressed in an obvious way in terms of r_0 by applying the condition that the functions be continuous at $r = r_0$. In § 3.14 of Volume 1 it is shown that $F^*/a = r_g^*$. Therefore, we finally obtain $r_g^* = a_{\max} \sin^3 r_0$, where r_0 is the boundary of the sphere. This last expression is the one presented earlier in the text.

II PHYSICAL PROCESSES IN THE HOT UNIVERSE

INTRODUCTION TO PART II

1 Introductory Remarks and Historical Review

In this, the second part of the book, we shall examine the physical processes which occur during the early epochs of the Universe under the assumption that the geometry and dynamics of the evolution are described by the homogeneous, isotropic model.

The physical processes in these early stages depend only on the composition of the matter which fills the Universe. These processes occur as the Universe evolves in accordance with a definite expansion law, examples of which were studied in detail in Part I of this book (to some degree, the processes themselves influence the expansion law). But even in the absence of such influence, the geometry and dynamics are by no means completely determined; they still depend on the composition of the matter.

In the following chapters, we shall present an account in accord with the matter composition that, with considerable certainty, follows from modern experimental data. The most important observational result relevant to a determination of the composition is that relic radiation—RR for short—exists. This leads to a picture of the Universe in which there are, on the average, about $10^{9\pm1}$ photons for each baryon. The latter consist of the protons and neutrons, free or bound in a nucleus. The value of this ratio depends on the density of matter and the Hubble constant. In general: $n_\gamma / n_B \approx 10^8/3\Omega_0 h_0^2$. The RR has a Planck spectrum with a temperature $T \approx 2.7$ K. Can this radiation be produced by astronomical objects? A simple analysis, which we will discuss later in this section, shows that the answer is no. The analysis shows that the radiation is always present in the Universe, back to the earliest stages of expansion. Knowledge of the ratio n_γ/n_B is sufficient for a characterization of the composition of the matter at each stage of the Universe. More detailed relations among the number of hydrogen nuclei, of helium nuclei, of photons, and of neutrinos are then consequences of the theory if one assumes the simplest possible conditions.

The point is that the presence of RR testifies to the high temperature of the matter at the beginning of the expansion (the "hot Universe"). At the high temperature and density that characterized the beginning of the expansion, certain elementary particles or nuclei are converted into others by processes which proceed rapidly. Therefore the numbers of particles of

different kinds are not independent and freely variable, but are specified by the laws governing these rapid processes. Only one number – the ratio n_γ/n_B of the number of photons to that of baryons—must be assigned.

More precisely, one says that in the early stages, thermodynamic equilibrium holds. For a specific description of this equilibrium state, one must know the specific entropy of the matter: with respect to the physical processes proceeding under equilibrium and adiabatic conditions, this entropy is conserved. The second thermodynamic variable, the density or specific volume of the matter, changes smoothly with time during the expansion. The corresponding time dependence can be considered known (see Part I). Chapter 6 is dedicated to a study of the theory of the equilibrium at high temperature.

With the lowering of the temperature that accompanies the expansion, the reactions proceed more slowly, until finally the rate of some process is comparable to the expansion. Thereupon the rate of the process is determined by the kinetics of the reactions. The most important process of this sort occurring during the early stages of expansion of the Universe is nucleosynthesis. In particular, given the prestellar matter present under cosmological conditions, one can expect 25–30% of the baryons to be converted into helium and 70–75% of the baryons to be left as protons. Such a result is not in contradiction with the observations, and thus increases confidence in the hot model. The problems of the kinetics and, in particular, of nucleosynthesis are discussed in chapter 7. Here, in a short introduction, the general situation is set forth very schematically and more simply. With respect to many questions of considerable importance, no clear approach exists. For example, do undiscovered particles of the neutrino type exist? Thus, in the detailed account below, a whole group of variants requires discussion. A comparison between the results and observations immediately leads to conclusions of interest for the theory of elementary particles.

In chapter 8 we examine processes occurring at a relatively late stage of expansion, when the only particles present are those that "survive" or are "preserved": photons, electrons, nuclei, and neutrinos (and hypothesized gravitons). Because of annihilation, antinucleons, positrons, and mesons disappear and (or) disintegrate, and nuclear reactions cease. At this stage, the interaction of photons with electrons, or, more specifically, the absorption, emission, and scattering of light, is most crucial.

Neutrinos and gravitons remain at this stage, but, regarding them, one can say that they are present "invisibly" and do not influence the photon processes. It is essential that here, in the second part of the book, the homogeneous, isotropic model be employed—isotropic with respect to the Friedmann expansion—without accounting for the spatial inhomogeneities that are assumed small at that epoch but will ultimately lead to the formation of galaxies (see the following part of the book for more about this).

The observed properties of the RR agree with those conclusions that follow from the evolution of the hot Universe; i.e., they support this picture. But let us pose a question which is more to the point: To what degree does the study of the RR confirm the theory of the hot Universe? Or, still more precisely: What part of the theory of the hot Universe can one consider proved? For the answers to these questions, the results of very important relevant research are set forth in chapter 8.

The interaction of RR with electrons would smear out any deviations in the spectral energy distribution of RR from that of the equilibrium spectrum predicted by the theory of the hot Universe. This is true if these deviations occur sufficiently early, for $z > 10^6$–10^5 or $t < 0.3$–30 yr from the beginning of the expansion. Deviations which arise later, connected with the processes of energy transfer to matter and with the consequent change in the RR spectrum, would be detected.

There have been attempts at an explanation of RR using the energy emanating from separate celestial sources (Parijsky 1968; Burbidge 1971; and Layzer 1968). Let us consider this suggestion briefly.

The integrated density of the equilibrium background radiation, $\epsilon_\gamma \approx 4 \times 10^{-13}$ ergs cm^{-3}, is almost 100 times greater than the integrated density of radiation from discrete sources: stars, galaxies, radio galaxies, and quasars. Note that the latter is a quantity calculated including the probable evolution of these sources. Moreover, the background radiation is, with great accuracy, highly isotropic. Now if one divides the energy density $\epsilon_\gamma \approx 4 \times 10^{-13}$ ergs cm^{-3} by the matter density 3×10^{-31} g cm^{-3} (the average density of matter in galaxies, according to the data of Oort 1958), one obtains $\approx 10^{18}$ ergs g^{-1} for the ratio. Thus, burning $\sim 20\%$ of all of the hydrogen would give enough energy to fill space with radiation at a temperature of ~ 3 K. However, we know that the burning of nuclear fuel in stars gives a spectrum very different from that of the radiation observed. In addition, the angular distribution of the stellar radiation so produced is discrete and not uniform. If the burning takes place in the distant past, then, in principle, the high-temperature radiation emitted by such sources can, because of the cosmological expansion, transform into low temperature radiation. As the expansion proceeds, however, the energy decreases, $\epsilon_\gamma = \epsilon_0(\rho/\rho_0)^{4/3}$, requiring more energy in the past than any reasonable limit.

Thus, one can attempt to explain the presently observed background as due to the liberation of energy in "ancient" stars, but this explanation requires the very improbable combination of many assumptions. One must assume a high, present-day average matter density $\rho_0 \gtrsim 2 \times 10^{-29}$ g cm^{-3}. It is moreover necessary to assume that practically all the nuclear energy of all the matter is liberated and that this energy is radiated by hypothetical "ancient stars" for a short period in the past ($t = 3 \times 10^6$ yr). Finally, one assumption must be made regarding the absorption and subsequent

reradiation of light by the dust that constitutes not less than 10^{-5} of all of the matter by mass. Such an explanation of the background seems totally unlikely (see Zel'dovich and Novikov 1967*a*).

A decisive experiment would be a measurement of the temperature of the neutrino background. In accordance with the predictions of the hot model, this temperature, at present, must be $T \approx 2$ K (see subsequent discussions) However, this "experiment of the century" presently lies far beyond the limits of possibility. Such an experiment, detecting neutrinos with $T \approx 2$ K, would prove not only that the hot model is correct, but also that the expansion proceeds isotropically from the very first seconds. The point is that, in the case of anisotropic deformation at the early stages of the cosmological expansion, the present-day spectrum of relic neutrinos can differ strongly from an equilibrium spectrum with $T \approx 2$ K (Doroshkevich, Zel'dovich, and Novikov 1967*a*, *b*, *c;* see also Part III).

Under the conditions of nuclear-energy liberation in the "ancient stars," the neutrinos generated would evidently not have an equilibrium spectrum, and "thermalization" of them would be impossible since matter is transparent to them. Their number density (particles cm^{-3}) must be of the order of the number density of nucleons, and their present-day energy, of the order of tens of keV. A determination of the number density of neutrinos would allow one to estimate the energy liberated. A determination of the present-day energy of the neutrinos would give the epoch in the past when the liberation of the energy occurs and would show whether or not one can explain the background in this fashion.

Attempts at an explanation of RR as the integrated radiation from radio sources (Parijsky 1968; Burbidge 1971) contradict the statistics of radio sources as well as the agreement of the RR spectrum with the Planck spectrum in the well-studied, long-wavelength region.

Thus, one can say that the general picture of the "hot Universe" is proved by observations directly for the period beginning several years from the singularity and lasting to the present day; i.e., it is proved for the period $10 \text{ yr} \lesssim t \lesssim 10^{10}$ yr.

For a discussion of the possibility of an anisotropic beginning for the expansion of the Universe and the distortion of the relic radiation connected with such anisotropy, see Part IV. Note too that the processes occurring in the period of the formation of galaxies, quasars, and so forth were not completely elucidated in the picture sketched here (for more about this subject, see Part III).

The classical picture set forth below is charge asymmetric, it being assumed that, in today's Universe, nature presents only baryons—matter—and practically no antibaryons—antimatter. Thus, it is assumed that distant galaxies and the intergalactic gas consist of matter.

From the point of view of the laws of physics as studied in the laboratory, i.e., of what is "natural," and from the point of view of "sound thinking,"

annot choose between the symmetric and asymmetric theories of the Universe. However, arguments supporting the asymmetric theory essentially rest on such astronomical data as the redshift and the relic radiation. In the literature one also meets the opposite point of view, according to which the Universe is charge-symmetric on the average. The different possible variants are based on different assumptions as to how distant from one another are regions with matter and regions with antimatter. Are there antimatter stars in our Galaxy? Are there galaxies and antigalaxies in a single cluster of galaxies? Below appear arguments supporting our view that the charge-symmetric model does not conform to reality. Nevertheless, there is no denying that the proponents and defenders of the symmetric theory (Klein and Alfvén) are logically consistent and clever in arguing the contrary.

The defenders of these or of other unorthodox concepts have thoroughly mastered the thesis that an attack is the best method of defense. Burbidge, Hoyle, and Alfvén emphasize the "exotic" and "astonishing" character and the low probability of the classical model with homogeneous and isotropic structure, with dynamic expansion from an initially singular state, and with asymmetric charge. This reproach would be valid if the theory of the hot Universe were an abstract work of "pure reason." But the whole point is that at the present time (it was not always so) this picture is necessary from an observational point of view! The inevitability of a strange world is just the formulation of the writer Danin, expressed on another occasion, but so accurately characterizing cosmology.

Now let us briefly discuss the historical development of those ideas which together constitute today's notion of the hot Universe. Proof that the Universe expands appeared roughly at the same time that modern ideas on nuclear physics arose. Thus, see the works of Friedmann in 1922 and 1924; the announcement in 1919 of the discovery by Rutherford of nuclear reactions; the works of Hubble in 1929; and the notice in 1932 describing the discovery of the neutron by Chadwick. Almost simultaneously there then arose, first, the questions of what the Universe and the matter filling it represent near the singularity and, second, what the physical preconditions for the answer to this question are. Already, the most general idea about the singularity, the infinite-density state at the beginning of the expansion, had led to the idea that the original, prestellar matter was more or less homogeneous—in stark contrast to the diverse structure and composition of the stars and of other celestial bodies today. Then, a pleasant-sounding but nebulous formulation appeared: the Universe is a single "primordial atom," or the entire Universe is a single nucleus or drop of nuclear fluid (Lemaître 1933).

The next stage, at which the idea of the hot Universe arose, is a remarkable example of the fruitful development of a theory in whose origin a chain of errors played a substantial role.

Gamow (1948) set himself the problem of providing a cosmological explanation for the abundances of the different nuclei and isotopes. The formulation of this problem was connected with the then-accepted but erroneous value of the Hubble parameter, $H_0 \approx 500\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$. This value corresponds to a time from the singularity of less than 2×10^9 yr. But the age of the Earth is estimated as 4×10^9 yr. Even allowing for errors in these quantities, one naturally assumed that the Earth as well as the Sun condense simultaneously from the primordial prestellar matter and that their chemical composition is determined by the composition of the prestellar matter. According to the hypothesis of Gamow, not only hydrogen and helium—the main constituents of stars and, in particular, of the Sun but also all the heavy nuclei right up to iron, lead, and uranium accumulate, in the course of the expansion, as products of reactions involving the prestellar, primordial matter.

A table of the abundances of the different elements and isotopes displays a general regularity: in nature, the isotopes with many neutrons predominate. From this, one draws the conclusion that free neutrons exist for a long time in the primordial matter. This requires a high temperature; hence the idea of a "hot Universe" appeared.

From today's point of view, one can point to a number of inconsistencies. First, the Hubble parameter has decreased by a factor of 5 to 10. Next, the Sun is a star of the second or third generation. Let us explain the significance of the latter point. The stars of the first generation form first. In massive stars of the first generation the nucleosynthesis of heavy nuclei (those heavier than He) occurs. With the explosions of these stars the heavy nuclei disperse into the interstellar gas from which the Sun and the Earth condense. It is true that, for the synthesis of heavy nuclei, free neutrons are necessary. However, these neutrons appear during slow, as well as during explosive, stages of the evolution of massive stars. On the other hand, notwithstanding the presence of free neutrons, the synthesis of heavy elements does not occur in the primordial, prestellar matter. There, nature is unsuccessful in surmounting the barrier at atomic weight five, for a nucleus of ^{4}He is unable to capture neutrons.

Nevertheless, both Gamow and later Alpher and Herman (1953) supposed that one would somehow manage to surmount the barrier at mass number $A = 5$ and obtain the intermediate and heavy elements of the periodic system through successive neutron addition and β-decay. The prevalence, in nature, of elements with a neutron excess was considered to be a confirmation of this hypothesis.

Since the isotopes survive over a long period of neutron capture, the conclusion was drawn that nucleosynthesis, the capture of neutrons, takes place at a high temperature. In this way an estimate of the radiation temperature of about 6 K was obtained!

For the complete logical development of the theory, an essential assumption was that the Universe is young. However, modern data bearing on the Hubble parameter allow a wholly sufficient time for the nucleosynthesis of the intermediate and heavy elements in stellar processes.

Thus the question about nucleosynthesis was stated improperly and, therefore, the answer to this question of the theory of the hot Universe does not coincide with observations. But theories have their destiny, and their development is not described by the straightforward scheme of "question and answer." An analysis of the conditions at high temperature at the beginning of the cosmological expansion led to the conclusion that there is a period, known as the radiation-dominated (RD) era, when the mass density of radiation is greater than the density of ordinary matter.

This unusual possibility follows from the laws of physics, and its unrealizability in the laboratory bewitched researchers. However, one can perhaps manage to achieve such conditions at the focus of a picosecond pulse of laser radiation. With the study of the properties, mechanics, and thermodynamics of a photon plasma, an RD state of matter, a new field began. The elucidation of this field and of the diverse consequences arising from the notion of the hot Universe acquired an interest of its own. The problem of nucleosynthesis at the beginning of the expansion was meanwhile solved by the labors of Hayashi (1950) and of Fermi and Turkevich (unpublished). The calculations predict the following composition of prestellar matter: hydrogen, 65–75%; ^{4}He, 25–35%. Other abundances are not required, for the remaining elements are synthesized in stars. Notice that the work of Fermi and Turkevich is not published; apparently, they used data about the reactions of isotopes of hydrogen which at that time could not be made public. Later, researchers considered the traces of other elements formed via cosmological nucleosynthesis: D, ^{3}He, ^{6}Li, and ^{7}Li.

Curiously, the originators of the hot model were mainly interested in the integral properties of the relic radiation—its density, pressure, and temperature—but not its spectrum. By means of not especially reliable considerations, Gamow in 1946 gave an estimate of the present-day temperature of 5–6 K.

Chiefly with a view toward self-criticism, we will remark on the alternative hypothesis of a "cold Universe." It was known (Peierls, Singwi, and Wroe 1952) that cold, superdense matter having the composition we encounter in stars (neutron and hyperon stars; see Vol. 1) gives, by means of nucleosynthesis occurring during expansion, only heavy nuclei. Hydrogen does not remain. Such a variant of the hypothesis of the "cold Universe" sharply contradicts the observed picture that hydrogen constitutes not less than 60% of all matter.

However, according to a remark by Zel'dovich (1962*b*), cold matter with abundant neutrinos transforms into pure hydrogen in the course of

the expansion. For this to occur in a mixture of protons, electrons, and neutrinos, it is only necessary that the Fermi energy of the cold neutrinos at all times exceed the Fermi energy of the electrons.

In connection with this hypothesis and also with the question of the average density of electromagnetic energy in the Universe, Doroshkevich and Novikov (1964) carried out—apparently first—a complete calculation of the broad spectrum of electromagnetic radiation in the evolving Universe, from the longest radio wavelengths to the ultraviolet. The Planck spectrum of radio radiation, which must be observed in the case of a hot Universe, was superposed on the curve obtained by adding together the contributions of radio sources and stars.

One notices from this work that the Planck spectrum is, in a definite radiation band, several orders of magnitude more intense than the sum of the spectra of the discrete sources. So the possibility of a decisive experiment, the observation of RR, became clear. Consider this as still another example of errors when these errors do not occur on purpose. As a poet once said about wineglasses, "One broken accidentally is a sign of good luck, while one broken on purpose bodes evil," i.e., you cannot force good luck. Indeed, the choice between the hot and cold model of the Universe depends on the results of this experiment. In the work of Doroshkevich and Novikov the two-humped curve first appeared, which for a long time has adorned the advertisements of the Bell system.

The subsequent history is well known. Penzias and Wilson (1965) stumbled across unexplained, regular radio noise at a wavelength of 7.3 cm. At the same time Dicke, Peebles, Roll, and Wilkinson (1965) were preparing an apparatus for the measurement of the radio background at a wavelength of 3 cm with the deliberate intention of verifying the theory of the hot Universe and determining the temperature of the relic radiation. At that time, working independently of Doroshkevich and Novikov, they evidently had made similar estimates. (See the book of Dicke 1970.) Then hearing of the results of Penzias and Wilson, Dicke and his group quickly interpreted their results as confirmation of the theory of the hot Universe and assigned a temperature of about 3 K.

It would be inappropriate to set forth the further development in the form of a historical review, as the succeeding results are destined for a systematic account. Here, in order to give a complete picture, we present only the final conclusions. The theory of the hot Universe, which constitutes the theory of a huge stage of the evolution of the Universe, is today finally established. The decisive factor is the existence and properties of RR. Those refinements which can follow (since the accuracy of all of the measurements conducted is not perfect) will not change the basic conclusions about the hot Universe. Refinements will rather give information about the details of processes that occur in the past.

Indeed, it is already clear that one will be able to (and will have to)

plain deviations within the bounds of the theory of the hot Universe. In particular, one will be able to account for small inhomogeneities, the formation of galaxies and quasars, and other analogous phenomena. Nucleosynthesis does not play a decisive role in the proof of the theory of the hot Universe. Nevertheless, that the theory predicts an abundance of ^{4}He close to that observed is doubtless an argument in its favor. Note, though, that the situation is not entirely clear, since the role of stars in the synthesis of He and of several other elements is still not precisely determined.

These remarks must not be considered merely dogmatic. Within the theory of the hot Universe many questions remain. As a measure of our understanding, we will discuss them in Parts II and III below. For a discussion of the deep, unsolved questions of principle that remain concerning the nature of the singularity, see Part IV.

5.2 Electromagnetic Radiation in the Universe—a Review of the Observations

A broad review of the observations of the entire spectrum of electromagnetic radiation in the Universe, from radio waves to gamma rays, serves as a good introduction to a detailed study of the RR that is most important in cosmology. Longair and Sunyaev (1971) recently gave a complete review of the electromagnetic radiation. We will make use of their figures and tables. We are interested in the average density of the radiation far from all discrete sources. Therefore, all quantities (the intensity F_ν, energy density ϵ_γ, and number density n_γ) are extrapolated to intergalactic space, with the contribution of our Galaxy subtracted from the observational data. While comparing Figure 27 and Table 2, one must remember the obvious relations

$$\epsilon_\gamma = \frac{4\pi}{c} \int F_\nu d\nu = \frac{4\pi}{c} \int \nu F_\nu d \ln \nu \,, \tag{5.2.1a}$$

$$n_\gamma = \frac{4\pi}{hc} \int \frac{F_\nu}{\nu} d\nu = \frac{4\pi}{hc} \int F_\nu d \ln \nu \,. \tag{5.2.1b}$$

The values given in the table are not equal to the area under the curve in Figure 27 because of the logarithmic scale in Figure 27. For each region of the spectrum, however, one has approximately

$$\epsilon_\gamma = \alpha \frac{4\pi}{c} (\nu F_\nu)_{\max} \,, \tag{5.2.2a}$$

$$n_\gamma = \beta \frac{4\pi}{hc} (F_\nu)_{\max} \,, \tag{5.2.2b}$$

where α and β are of order unity. By astronomical tradition, the frequency $\nu = c/\lambda$ is used. We thus have Planck's constant $h = 6.54 \times 10^{-27}$ erg-sec and not $\hbar = h/2\pi$.

Now let us give a short description of the different parts of the spectrum

It is assumed that the behavior of the left side of the curve in Figure 27 is dominated by discrete radio sources. Its fundamental part is given by the power law $F_\nu = A\nu^{-0.75}$. Taking into consideration the Rayleigh-Jean formula $F_\nu = 2kT/\lambda^2$, one can present these data by means of the formula for the effective temperature,

$$T_{\text{eff}} \approx 3 \times 10^{-5}\lambda_{\text{cm}}^{2.7}\,, \qquad (5.2.3$$

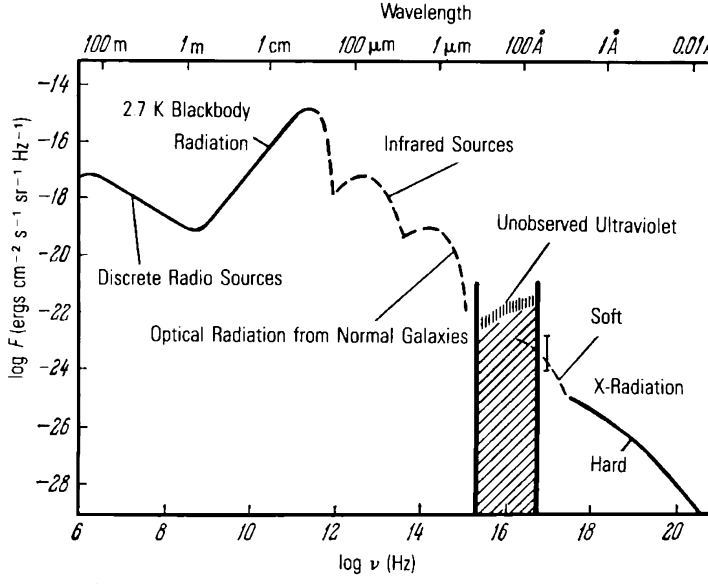


FIG. 27.—The spectrum of electromagnetic radiation in the Universe

TABLE 2

ENERGY DENSITY AND NUMBER OF PHOTONS OF THE BACKGROUND RADIATION IN VARIOUS WAVELENGTH BANDS

Band	Energy Density of the Radiation (ergs cm^{-3})	Number Density of Photons (cm^{-3})
1a. Long wave radio	$\sim 10^{-7}$	~ 1
1b. Relic radio	0.25	400
2. Infrared	$\sim 10^{-2}$	~ 1
3. Optical	$\sim 3\times 10^{-3}$	$\sim 10^{-3}$
4a. Soft X-ray ($E<1$ keV)	$10^{-4}-10^{-5}$	$3\times 10^{-7}-10^{-8}$
4b. Hard X-ray ($E>1$ keV)	10^{-4}	3×10^{-9}
5a. Soft γ-ray ($E\approx 1$–6 MeV)	$\lesssim 10^{-5}$	$\lesssim 3\times 10^{-12}$
5b. Hard γ-ray ($E>10$ MeV)	$<10^{-5}$	$<10^{-12}$

giving the contribution from various sources to the total T_{eff}. For $\lambda =$ [illegible] cm, ν [illegible] 178 MHz and $T_{\text{eff}} \approx 30$ K (see Bridle 1967, 1968; Reber 1968). At this wavelength, our Galaxy gives a minimum temperature $T_{\text{eff}} \approx 80$ K. In the direction perpendicular to the plane of the Galaxy, this contribution is removed by means of measurements and calculations of the angular dependence of the Galactic contribution. What is the origin of the extragalactic part of this radio radiation, a part that causes a temperature T_{eff} [illegible] 30 K? Discrete sources give more than half of it, but very few (5%) of these sources are identified with optical sources. The rest is ascribed to discrete sources that are too faint to be seen by present telescopes.

Notice the kink at long wavelengths ($\lambda > 100$ m). Since the ionosphere surrounding the Earth reflects long-wavelength radio waves, measurements at such wavelengths are carried out with the help of satellites and interplanetary probes up to a wavelength of $\lambda \approx 10^5$ cm or a frequency of ν [illegible] 0.1 MHz (see, for example, Alexander *et al.* 1969). A distinct maximum appears at 4 MHz: $F_{\max} \approx 10^{-17}$ ergs $(\text{cm}^2\,\text{s}\,\text{sr}\,\text{Hz})^{-1}$.

In the next frequency band, 10^9–10^{12} Hz, RR dominates. It strongly exceeds the total of all other types of radiation with respect to n_γ as well as with respect to ϵ_γ (see Table 2). We will discuss it separately in the next section.

The infrared part of the spectrum, 10^{12}–10^{14} Hz, is poorly known today. This part is shown by the dashed line. The cited estimates are taken from the work of Low and Tucker (1968). The nucleus or central part of our Galaxy possesses a high infrared luminosity $L \approx 10^8 L_\odot$ for $\lambda \approx 100\,\mu$. Recent theories explain this as arising from the dust surrounding the nucleus: Optical radiation is absorbed, the dust heats up, and radiates at T [illegible] 1000 K. The luminosity of quasars also often has a maximum in the infrared region. As Shklovsky indicated, however, it is possible to explain infrared radiation from quasars and the nuclei of galaxies by means of synchrotron radiation from especially small regions with strong magnetic fields. Also, an interesting hypothesis developed by Bisnovatyi-Kogan and Sunyaev (1971) connects the infrared maximum with a collective mechanism in the plasma near relativistic objects.

Measurements in the optical region are difficult because of the background created by the scattered light of the Sun. The observations of Roach and Smith (1968) present an upper limit to the extralactic radiation.

Many theorists have calculated the spectrum of radiation from galaxies assuming a constant number of galaxies in a given comoving volume or by making various assumptions concerning evolution. The results are weakly sensitive to the cosmological models (see, for example, Doroshkevich and Novikov 1964). Since the brightest galaxies are observed up to redshifts $z \lesssim 0.5$, the calculations are based on an extrapolation into the past. Now, the majority of stars have a mass $M \lesssim M_\odot$. They can shine without

noticeable evolution for more than 10^{10} years while remaining on the main sequence; hence, the disappearance of galaxies is improbable. Thus one must assume that the density of galaxies in the past is not greater than now. However, it is possible that in young galaxies the percentage of very bright stars is greater than now. Having mentioned this proviso, we give data in Table 2 taken from the work of Partridge and Peebles (1967*a*, *b*). The observable upper limit is several times greater.

We must point out that the energy density in the optical band calculated above constitutes a very small part of the available nuclear energy. Even taking the lower limit of 3×10^{-31} g cm^{-3} as an estimate of the mass density, we would obtain $\epsilon_{\text{nuc}} \approx 0.5$ eV cm^{-3} or 2.7×10^{18} ergs g^{-1} upon the conversion of 30% of the hydrogen into helium. This exceeds, by a factor of 200, the energy density of the optical background. Therefore, no difficulties arise in explaining the optical background as the total radiation of all of the galaxies. Direct observations of the optical background are important. However, a measurement of the brightness of the night sky is difficult since zodiacal light and light from the stars of our own Galaxy interfere. The latest measurements give a flux $F_\nu < 10^{-19}$ ergs (cm^2 s sr Hz)$^{-1}$ at a wavelength of 5300 Å (Roach and Smith 1968) and $F_\nu < 10^{-20}$ ergs (cm^2 s sr Hz)$^{-1}$ at 4100 Å (Lillie 1969). These measurements correspond to an energy density $\epsilon_\gamma < 10^{-2}$ eV cm^{-3}.

For ultraviolet radiation, direct measurements from rockets (Kurt and Sunyaev 1967*a*, *b*) give an upper limit $F_\nu < 3 \times 10^{-21}$ ergs (cm^2 s sr Hz)$^{-1}$ for 1350 Å $> \lambda >$ 1216 Å. Note that the diagram includes contributions on both sides of the Lyman-α (Lα) line (1350 Å $> \lambda >$ 1225 Å; 1180 Å $> \lambda >$ 1050 Å), which has a wavelength of 1216 Å. The Lα line itself predominates in the radiation from the gas in our Galaxy—including the region near the Sun. Of great interest is the data for $\lambda >$ 1216 Å. Why? Intergalactic gas must radiate Lα. Displaced in wavelength by the cosmological redshift, the observed radiation falls in the interval 1340 Å $> \lambda >$ 1225 Å. The cited measurements thus give an estimate for the upper limit on the density of the intergalactic gas: $\Omega_{\text{gas}} = \rho_{\text{gas}}/\rho_c < 4$. Evidently, more precise measurements are necessary.

Radiation whose wavelength is shorter than that of ultraviolet radiation ($\lambda <$ 916 Å, the Lyman continuum) is strongly absorbed by neutral hydrogen in the Galaxy, and up to now no method for its direct measurement exists.

Sunyaev (1969*a*, *b*) proposed a clever way of estimating the amount of short-wavelength radiation with wavelengths less than that of the Lyman continuum. The method entails observations of the rarefied, neutral-hydrogen halo of galaxies and of the bridges between galaxies, the presence of neutral hydrogen being detected by observations of the 21 cm line. The presence of low-density neutral hydrogen around galaxies is not compatible with a large ultraviolet flux. This consideration gives the upper limit on

du[illegible] in the ultraviolet region that is shown in Figure 27. From this it follows that the heated intergalactic gas producing an ultraviolet flux cannot exist in excessively large quantities; i.e., its density must correspond to the estimate $\Omega_{gas} < 0.3$. (This value of Ω_0 relates to the situation when $H_0 = 75$ km s^{-1} Mpc^{-1}. For $H_0 \lesssim 50$ km s^{-1} Mpc^{-1}, the result is less definite.)

Beginning with $\lambda < 70$ Å, the Galaxy becomes sufficiently transparent that short wave X-rays and γ-rays are measured directly from rockets and satellites (see the review of Longair and Sunyaev 1971).

If one excludes about 30 known discrete X-ray sources, the background is isotropic, within 10%. Apparently, this radiation is not explained as the thermal radiation of a hot intergalactic gas since the spectrum in the hard region, $E > 1$ keV, is described by a power law $F_\nu \sim \nu^{-\alpha}$ and not by an exponential law $F_\nu \sim e^{-\beta\nu}$. There is some basis to suppose that the X-rays and γ rays or a part thereof result from inverse Compton scattering of low-energy photons by relativistic electrons (Felten and Morrison 1966). An unanswered question then remains. Are these electrons distributed more or less homogeneously in space or, as Longair and Sunyaev (1969*a*, *b*) assume, do they produce X-rays and γ-rays right in the neighborhood of a source of electrons? In the latter case, with improved angular resolution, the quasi isotropic X-ray background will appear to arise from many discrete sources. A detailed discussion relates more to the physics of cosmic rays than it does to cosmology, so we shall now return to the contribution to the electromagnetic spectrum which is most important for cosmology– RR.

5.3 Observational Proof of the Existence of the Relic Radiation (RR)

In Figure 28 and in Table 3 are summarized the majority of RR measurements through 1972. Many of the data in Table 3 are from Peebles (1971*a*). They are arranged in the table in order of decreasing wavelength.

The so-called antenna temperature T_a, ordinarily used by radio astronomers appears in the third column. This temperature is calculated from the observed flux F_λ at a given wavelength according to the Rayleigh-Jeans formula

$$F_\lambda = \frac{2kT_a}{\lambda^2} = 2kT_a\left(\frac{\nu}{c}\right)^2. \tag{5.3.1}$$

If the flux of radiation is due to a thermal source, then T_a corresponds to the true temperature of the radiation only when $h\nu/kT \ll 1$. In the fourth column the thermodynamic temperature, calculated according to the Planck formula,

$$F_\nu = \frac{2h\nu^3}{c^2}\left[\exp\left(\frac{h\nu}{kT}\right) - 1\right]^{-1}, \tag{5.3.2}$$

is presented.

With the exception of numbers 16–18, all of the measurements were

TABLE 3

OBSERVATIONS OF THE RELIC RADIATION

Number	Wavelength (cm)	Antenna Temperature (K)	Thermodynamic Temperature (K)	Investigators
1.....	73.5 49	3.7 ± 1.2	3.7 ± 1.2	Howell and Shakeshaft 1967
2.....	21.2	3.2 ± 1.0	3.2 ± 1.0	Penzias and Wilson 1967
3.....	20.9	2.5 ± 0.3	2.5 ± 0.3	Pelyushenko and Stankevich 1969
4.....	20.7	2.8 ± 0.6	2.8 ± 0.6	Howell and Shakeshaft 1966
5.....	7.35	3.0 ± 1.0	3.1 ± 1.0	Penzias and Wilson 1965
6.....	3.2	2.8 ± 0.5	3.0 ± 0.5	Roll and Wilkinson 1966
7.....	3.2	$2.47^{+0.16}_{-0.21}$	$2.69^{+0.16}_{-0.21}$	Stokes, Partridge, and Wilkinson 1967
8.....	1.58	$2.35^{+0.12}_{-0.17}$	$2.78^{+0.12}_{-0.17}$	Stokes, Partridge, and Wilkinson 1967
9.....	1.50	1.56 ± 0.8	2.0 ± 0.8	Welch, Keachie, Thornton, and Wrixon 1967
10.....	0.924	2.44 ± 0.26	3.16 ± 0.26	Ewing, Burke, and Staelin 1967
11.....	0.856	$1.81^{+0.16}_{-0.21}$	$2.56^{+0.17}_{-0.22}$	Wilkinson 1967*a*
12.....	0.82	2.1 ± 0.7	2.9 ± 0.7	Puzanov, Salmonovich, and Stankevich 1967
13.....	0.358			Kislyakov *et al.* 1971
14.....	0.33	0.89 ± 0.32	$2.46^{+0.40}_{-0.44}$	Boynton, Stokes, and Wilkinson 1968
15.....	0.33	1.01 ± 0.20	2.61 ± 0.25	Millea, McColl, Pederson, and Vernon 1971
16.....	0.264	1.2 ± 0.4	3.2 ± 0.5	Field and Hitchcock 1966*b*
17.....	0.264	1.6 ± 0.6	3.7 ± 0.7	Peimbert 1968
18.....	0.264	0.93 ± 0.11	2.83 ± 0.15	Bortolot, Clauser, and Thaddeus 1969

carried out with radio astronomical instruments. The observations consist of a measurement of the receiver noise due to the RR flux. All of the internal receiver noise, along with noise from the atmosphere and the Galaxy, is interference. The apparatus is calibrated with the help of an artificial blackbody cavity maintained at the temperature of liquid helium and located at the opening of the antenna.

Galactic noise and the background from discrete radio sources have falling spectra with different spectral indices. For $\lambda > 21$ cm, therefore, the corrections due to them are large. The respective points showing the flux from RR and from discrete sources appear on the left side of Figure 28. It is difficult to extend the measurements to wavelengths $\lambda > 80$ cm.

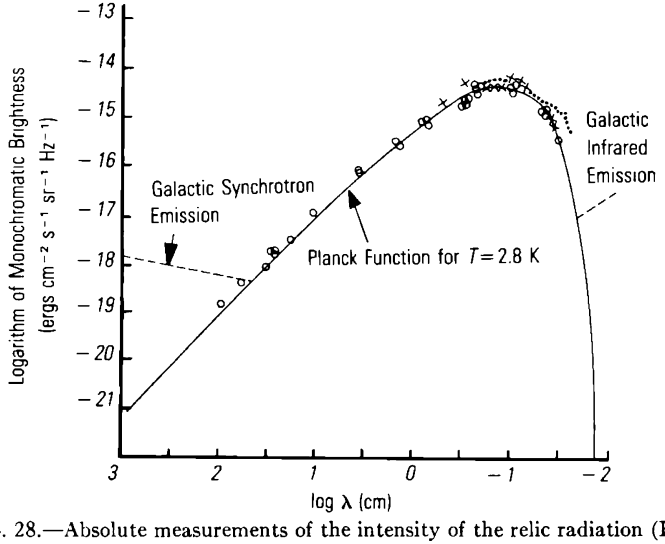


FIG. 28.—Absolute measurements of the intensity of the relic radiation (RR)

At short wavelengths the atmospheric noise is strong; for $\lambda \approx 0.33$ cm it reaches 10–12 K. For still shorter wavelengths the receiver must be transported beyond the atmosphere. Measurements 16–18 of Table 3 were done by means of an indirect spectroscopic method requiring optical observations of the lines of interstellar CN. The theory behind the interpretation of these observations is as follows. Molecules in interstellar space with low-lying excitation levels (we will denote the excitation energy by E_1) are excited by RR with a frequency $\nu_1 = E_1/h$. In the absence of collisions, the ratio of the equilibrium concentration of excited molecules to that of molecules in the ground state is

$$n_1/n_0 = \exp(-E_1/kT),$$

where T is the effective temperature of the radiation that excites the molecules. The concentrations n_1 and n_0 are determined according to observations of the optical absorption lines of these molecules.

Interestingly, McKellar (1941) noticed many years ago that the molecule CN is observed in the interstellar gas not only in the ground state but also in the rotational excited state that corresponds to an excitation temperature of ~2.3 K. McKellar arrived at this conclusion by studying the absorption lines of gas along the line of sight to the star ζ Ophuchi. The explanation of this effect remained unclear. Only after the discovery of RR did Wolf (1967) and Field and Hitchcock (1966*a*, *b*) explain this effect as the excitation of the interstellar molecule by RR. The analogous effect was thereupon observed in the lines of the gas near the star ξ Persei. Thaddeus and coworkers (1969; number 18 of Table 3) then observed the effect in the spectra of 11 stars in different directions on the celestial sphere. This proves that the excitation mechanism acts everywhere in the Galaxy, thus strengthening the notion that the excitation is due to RR. For the molecules CH and CH^+, and also for other levels of CN, the corresponding excited states are not found, although the transitions to the ground states of these molecules are observed. Therefore, one can only give an upper limit to the temperature. Corresponding limits are presented in Figure 28 (Bortolot, Clauser, and Thaddeus 1969).

Data on the intensity of electromagnetic radiation for wavelengths around 1 mm were obtained (Shivanandan, Houck, and Harwit 1968; Houck and Harwit 1969*a*, *b*) with the help of semiconducting bolometers on rockets ascending to an altitude of more than 100 km. The receivers were sensitive to the wide wavelength band from 0.4 to 1.3 mm. Measurements done over the course of several flights gave a sensational result—the average flux in this region of wavelengths corresponds to an energy density of order 10–20 eV cm^{-3}. This is significantly greater than a thermal flux whose temperature is $T = 2.7$ K. Furthermore, if this flux is distributed continuously throughout the wavelength interval and is incident on us from outer space (i.e., does not arise in the upper atmosphere or in the solar system), then the measurements contradict the upper limits on the flux in this region that were obtained by the interstellar molecule method. Moreover, measurements did not exhibit any anisotropy. One can obtain agreement with the molecular measurements by assuming that the excess radiation is concentrated in narrow lines not coinciding with the lines of the molecules. In subsequent publications of this group, ϵ_γ has varied but remains above its equilibrium value. Muehlner and Weiss (1970) have found that the total flux is many times larger than the equilibrium value.

Balloon experiments in one of the spectral intervals drew much attention to this question. If they were correct, the theory of the hot Universe would have to be modified substantially. However, with an apparatus cooled to superfluid helium temperature, Blair *et al.* (1971) of the Los Alamos Labo-

[illegible] conducted more careful measurements. The flux they observed corresponds to a temperature T 3.1 ($+0.5$, -2.0) K. Then, Muehlner and Weiss (1973) followed with a new series of balloon experiments employing improved apparatus. They obtained the result $T \approx 3$ K. Finally, an announcement has been made concerning new results of another Los Alamos group. These measurements, in the infrared part of the spectrum, agree with a Planck spectrum with $T = 2.7$ K. The conclusion is that the other measurements, which gave $\epsilon_\gamma > aT_{RJ}^4$, have been completely refuted.

The most precise measurements in fact do not contradict the relation $\epsilon = aT_{RJ}^4$, and are thus consistent with a thermal spectrum for RR. This result is important for the theory of the hot Universe. It indicates the small level of deviations from the idealized Friedmann picture and also suggests a primordial origin for the radiation.

The quadratic dependence of F_ν on ν compatible with the Rayleigh-Jeans law manifests itself in the constant value of the antenna temperature T_a, presented in Table 3, right up to $\lambda \approx 0.8$ cm. But this one fact still does not demonstrate the existence of an equilibrium Planck spectrum. Why? One can conceive of a spectrum,

$$F = \frac{1}{\alpha} F_{th}(\alpha T) ,$$

with the much higher temperature $\alpha \times 2.7$ K, $\alpha > 1$, satisfying the same relation in the Rayleigh-Jeans region. Indeed, processes exist (see the next chapter) that distort the Planck spectrum while preserving the $F \sim \nu^2$ law in the long-wavelength region.

Therefore, the observations at $\lambda \approx 0.33$ cm and $\lambda \approx 0.264$ cm (numbers 11 18 in Table 3) elicit great interest because, for $h\nu/kT \approx 2$, the Planck law deviates markedly from the extrapolated Rayleigh-Jeans law. Thus, $F_{Pl}/F_{RJ} = (e^2 - 1)^{-1} 2^{-1} = 0.315$ for $h\nu/kT = 2$. Those who published these measurements maintain that the measurements are not compatible with a power law, but agree well with a thermal spectrum.

There is further evidence that the radiation in the band from 50 cm to 0.3 cm does not arise from discrete sources but is primordial or relic in nature. The evidence consists of the isotropy of the radiation, and of the absence of any time or angular variations ($\Delta T/T < 10^{-3}$). This question is considered in more detail in the discussion of the degree of homogeneity and isotropy of the Universe, a discussion that appears in chapter 16.

It would be most desirable to increase the accuracy of the experiments that determine the density and the spectrum of the radiation. One can relate small deviations from a Planck spectrum to thermal processes in the Universe, for there will be perturbations from which galaxies form. The difficulty in performing the experiments is counterbalanced by their theoretical significance.

6 THERMODYNAMIC EQUILIBRIUM AT THE BEGINNING OF THE COSMOLOGICAL EXPANSION

6.1 Basic Epochs in the Evolution of the Hot Universe

Before proceeding with a detailed examination of the processes that occur in the course of the expansion of the hot Universe, we shall attempt to sketch a general picture of the evolution.

As we saw in § 5.3, the total energy density of RR today is $\epsilon_\gamma \approx 4 \times 10^{-13}$ ergs cm^{-3}, which corresponds to a mass density $\rho_\gamma \approx 4 \times 10^{-34}$ g cm^{-3}. Such a density is much less than the average density of ordinary matter (the matter in galaxies averaged over the total volume) which, as we saw above, is not less than $\rho_{\text{Galaxy}} \approx 10^{-31}$ g cm^{-3}. Thus $\rho_\gamma \ll \rho_{\text{Gal}}$ at present, so the gravitational force due to ρ_γ plays no role at all in the present-day dynamics of the expansion. Today, the expansion law is determined by ordinary matter with vanishing pressure, $P = 0$. Hence, as is shown in Part I, the scale varies according to the law

$$a \sim t^{2/3} . \tag{6.1.1}$$

The density of ordinary matter varies as

$$\rho_m \sim a^{-3} . \tag{6.1.2}$$

In contrast, although the RR photon number density varies in the same way (i.e., $n_\gamma \sim a^{-3}$), the energy of each photon decreases ($E_\gamma \sim a^{-1}$), and so the RR energy density varies as

$$\epsilon_\gamma \sim E_\gamma n_\gamma \sim a^{-4} , \qquad \rho_\gamma = \epsilon_\gamma / c^2 \sim a^{-4} . \tag{6.1.3}$$

From equations (6.1.2) and (6.1.3) it follows that

$$\frac{\rho_m}{\rho_\gamma} \sim a . \tag{6.1.4}$$

Thus, in the past, there was an epoch when $\rho_m \approx \rho_\gamma$. At that time the scale factor was a_{eq} and the redshift z_{eq}, where

$$1 + z_{\text{eq}} = \frac{a_0}{a_{\text{eq}}} = \left(\frac{\rho_m}{\rho_\gamma}\right)_0 \approx 4.5 \times 10^4 \Omega_0 h_0^2 . \tag{6.1.5}$$

For still earlier epochs ($z > z_{\text{eq}}$, $a < a_{\text{eq}}$)

$$\rho_\gamma > \rho_m . \tag{6.1.6}$$

These early epochs, during which condition (6.1.6) is fulfilled, are called the "radiation dominated" (RD) era. During the RD era, the evolution is determined by the gravitational field of an ultrarelativistic gas with $P = \epsilon/3$ and has the form (chap. 1) $a \sim t^{1/2}$. In agreement with the results of § 1.5 and § 1.2, the radiation temperature at all stages of the expansion varies as

$$T_\gamma \sim a^{-1}. \tag{6.1.7}$$

Consequently, the temperature is high during the early stages of the expansion. Indeed, both high temperatures and high matter densities prevail during the RD stage, though $\rho_\gamma \gg \rho_m$. Under these conditions, the ordinary matter exists as a plasma in a state of complete ionization and is in thermodynamic equilibrium with the radiation. This is all described in more detail in the following sections and chapters.

We emphasize that the ratio n_γ/n_m hardly changes at all during the entire temporal course of the expansion. At a late stage this constancy of n_γ/n_m is guaranteed by the conservation of RR photons, while at an early stage it is guaranteed by the condition of thermodynamic equilibrium.

When the temperature in the plasma is sufficiently high, i.e., when the time elapsed from the beginning of the expansion is small, pairs of particles and antiparticles are also found in equilibrium. For the appearance of a significant number of such pairs, it is necessary that the condition $kT \gtrsim mc^2$ be fulfilled where m is the rest mass of the particle. For example, there are many electron-positron pairs when $T \gtrsim 5 \times 10^9$ K. For still earlier times and correspondingly higher temperatures, pairs of still heavier particles and antiparticles are found in equilibrium.

The thermodynamics of matter at high temperatures is set forth in detail in chapters 7 and 8 of Volume 1, *Stars and Relativity*. Of course, no complicated nuclei can exist under these conditions. They would be split within an instant. Rather, nucleons exist in the form of neutrons and protons (and also in the form of still shorter-lived particles and "resonances"), constantly transforming into each other. Assuming that the condition $kT > mc^2$ is fulfilled, the number of particles and antiparticles of each kind is roughly equal to the number of photons. Small differences are caused by statistical factors characterizing the kind of particle. Thus, for each nucleon present in the Universe today, $\sim 10^8$ nucleon-antinucleon pairs are present in the earliest epoch of the expansion, when $T > 10^{13}$ K. From this point of view, the presently existing nucleons are the result of a small excess in the number of nucleons (10^{-8} of this) over that of antinucleons during that early stage. It is just this surprisingly small excess, whose average is not strictly equal to zero, which Burbidge and others emphasize when speaking about the "strangeness" of the charge-asymmetric model.

Having made these openings remarks, we now proceed to divide the expansion of the Universe into the following epochs, which we merely enumerate here but shall describe in detail later:

1. The hadron era, which occurs when $t \lesssim 10^{-6}$ s and $T > 10^{13}$ K. In addition to the neutrinos, mesons, and electron-positron pairs in equilibrium in the plasma, there are many nucleon-antinucleon pairs. The physics of this era is examined further in §§ 6.3–5.

2. The lepton era, which occurs when $10^{-6}\text{ s} < t < 10\text{ s}$ and $10^{13}\text{ K} > T > 5 \times 10^{9}$ K. Nucleon-antinucleon pairs annihilate by this time. (The peculiar viewpoint of Omnès is analyzed in § 6.3.) Photons, neutrinos, electron-positron pairs, and a small admixture of nucleons remain in equilibrium in the plasma. At the beginning of this stage there are also mesons, but they subsequently decay. Toward the middle of the era, processes involving neutrinos become slow with respect to the expansion. The neutrinos cease to interact with the plasma. They move freely, cooling as a result of the expansion, and remain noninteracting throughout all of the following epochs up to the present. At the end of this period the electron-positron pairs annihilate, and nucleosynthesis occurs. The processes occurring at this time are described in chapter 7.

3. The photon-plasma era, which occurs when $10\text{ s} < t < 10^{12}$ and $5 \times 10^{9}\text{ K} > T > 10^{4}$ K. Plasma and radiation are in equilibrium during this period, while the neutrinos are free. See chapter 8 for a detailed description.

4. The era after the hydrogen in the primordial plasma recombines. Somewhat later than $t \gtrsim 10^{12}$ s, when the temperature falls to a few thousand degrees, the primordial plasma recombines and becomes transparent to the relic radiation. After this the formation of separate heavenly bodies becomes possible by means of the gravitational instability of the originally small density fluctuations. This period, and other questions connected with the problem of the formation of heavenly bodies, are considered in Part III of the book.

Hypothesized gravitons are probably present at all epochs, but for all practical purposes they do not interact with the remaining particles during any stage after the "Planck" time $t_{\text{Pl}} \approx 10^{-43}$ s.

6.2 The Cosmological Expansion of the High-Temperature Plasma and the Condition of Thermodynamic Equilibrium

For an extended period in the evolution of the hot model, the density of nucleons and electrons is small compared with the density of photons and other massless particles. Electron-positron pairs are numerous only when the temperature is greater than the electron rest mass, $kT \gtrsim m_e c^2$ or $T \gtrsim 5 \times 10^{9}$ K. If this inequality is fulfilled, one can regard electrons and positrons as relativistic particles. The same assertion applies to still heavier

particles at correspondingly greater temperatures. Consequently, the following relation holds approximately at a sufficiently early stage:

$$P = \tfrac{1}{3}\epsilon = \tfrac{1}{3}\rho c^2 , \tag{6.2.1}$$

where P is the pressure, ϵ the energy density, and ρ the mass density. We shall assume that the mass spectrum is bounded, so that there is always a temperature such that $kT > M_m c^2$, where M_m is the maximum possible particle mass. We also assume that for this temperature, the interactions among the particles are not important. For a contrasting point of view, that of the theory of Hagedorn, see § 6.4.

For the equation of state described by (6.2.1), the expansion law was deduced in § 1.8. This law has the form

$$\begin{aligned} \rho &= \frac{3}{32\pi G t^2} = \frac{4.5 \times 10^5}{t^2}\ \mathrm{g\ cm^{-3}} , \\ \epsilon &= \rho c^2 = \frac{4 \times 10^{26}}{t^2}\ \mathrm{ergs\ cm^{-3}} . \end{aligned} \tag{6.2.2}$$

If the gas consists only of photons, then we would have

$$\epsilon = \sigma T^4 , \quad \text{where} \quad \sigma = \frac{\pi^2 k^4}{15\hbar^3 c^3} = 7.57 \times 10^{-15}\ \mathrm{ergs\ cm^{-3}\ K^{-4}} . \tag{6.2.3}$$

Taking account of the various kinds of particles that are in equilibrium, we may write instead

$$\epsilon = \rho c^2 = \kappa \sigma T^4 , \tag{6.2.4}$$

where κ is a dimensionless coefficient greater than unity. We then obtain

$$T(\mathrm{K}) = 1.5 \times 10^{10}[t(\mathrm{s})]^{-1/2}\kappa^{-1/4} , \qquad t(\mathrm{s}) = \frac{2.3 \times 10^{20}}{[T(\mathrm{K})]^2\kappa^{1/2}} , \tag{6.2.5}$$

$$T(\mathrm{MeV}) = 1.3[t(\mathrm{s})]^{-1/2}\kappa^{-1/4} , \qquad t(\mathrm{s}) = \frac{1.7}{[T(\mathrm{MeV})]^2\kappa^{1/2}} . \tag{6.2.6}$$

Finally, keeping in mind that the average energy of a particle is of order $3kT$, we obtain for the density of all kinds of particles

$$n \approx \frac{\epsilon}{3kT} = 6.4 \times 10^{31}\kappa^{1/4} t^{-3/2}\ \mathrm{cm^{-3}} . \tag{6.2.7}$$

As was already noted, all particles are in thermodynamic equilibrium at high temperature. Indeed, for the existence of thermodynamic equilibrium, it is necessary that the rates of the processes establishing equilibrium proceed faster than the expansion rate. Speaking more precisely, it is necessary that the duration of a process that tends to establish equilibrium be much less than the characteristic time for a change in the parameters of the plasma such as ρ and T.

In an isotropic solution, $\rho = \alpha(Gt^2)^{-1}$, where α is of order unity. Therefore, the time required to change the density from ρ to $e^{-1}\rho \approx 0.4\rho$, is of order $\Delta t = (G\rho)^{-1/2}$. Thus, Δt is of order t. On the other hand, the time required for the establishment of equilibrium is

$$\tau = (\sigma n v)^{-1} . \tag{6.2.8}$$

Here σ is the reaction cross section, n is the density of particles, and v is their velocity. At high temperatures, $v \approx c$. The quantity n is determined according to formula (6.2.7): $n = n_1 t^{-3/2}$ with n_1 a constant. Therefore

$$\tau = t^{3/2}(\sigma n_1 c)^{-1} . \tag{6.2.9}$$

For thermodynamic equilibrium it is thus necessary that

$$\tau = t^{3/2}(\sigma n_1 c)^{-1} < \Delta t \approx t , \tag{6.2.10}$$

or

$$\sigma > t^{1/2} n_1 c . \tag{6.2.11}$$

Therefore thermodynamic equilibrium occurs for $t \to 0$, if σ does not decrease sufficiently quickly as the energy of the particles grows. In general, one can only hope that condition (6.2.11) is fulfilled, though, for example, there is no doubt that the number of electron-positron pairs at high temperatures does not differ from the equilibrium value. To illustrate, let us consider the instant when $T \approx 1$ MeV, $t \approx 1$ s, and $n_{e^+} \approx n_{e^-} \approx 10^{31}$ cm^{-3}. The annihilation cross section is of order 10^{-24} cm^2, and the velocity of the particles is of the order of the velocity of light. The time for the establishment of equilibrium is consequently of order

$$\tau \approx \frac{(1\text{ s})^{3/2}}{\sigma n_1 c} \approx 10^{-17}\text{ s} .$$

Since τ is negligibly small in comparison with $t \approx 1$ s, total equilibrium with respect to the reactions $e^+ + e^- \leftrightarrow 2\gamma$ is assured. Precisely the same result holds for the establishment of equilibrium for muon pairs $\mu^+\mu^-$ and also for meson and baryon pairs of all kinds, though at correspondingly much higher temperatures. Only at the singularity itself, when $t \leq (\hbar c/e^2)t_{\text{Pl}}$ [recall that $t_{\text{Pl}} = (G\hbar/c^5)^{1/2} \approx 10^{-43}$ s], can the condition for the establishment of equilibrium not be fulfilled for charged particles and photons. Inasmuch as the equilibrium condition is fulfilled later with much room to spare, one cannot expect noticeable deviations from equilibrium.

The situation for gravitons is not clear (see § 7.2).

More will be said in chapter 7 about the boundaries of the epochs where the condition of thermodynamic equilibrium is fulfilled for one particle or another.

In subsequent sections of this chapter we will examine the hadron stage of the expansion of the Universe. This is the stage when the temperature is higher than the rest-mass energy of nucleons and when many nucleon-antinucleon pairs exist in equilibrium.

6.3 The Hadron Era in the Evolution of the Universe

For temperatures comparable to or higher than the rest-mass energy of the proton and the neutron, the number of nucleons and antinucleons in the plasma becomes of the order of the number of photons. One can find the corresponding, exact formulae in Landau and Lifshitz (1964). The difference between the number of nucleons and the number of antinucleons is equal to the number of nucleons observed today; i.e., this difference is about 10^{-9} of the number of photons. This small excess of nucleons is specified as an initial condition, so that upon expansion, the cooling of the plasma and the annihilation of the pairs gives the picture of the Universe observed today. At high temperatures, this difference is small with respect to the total number of nucleons and antinucleons, so that with sufficient accuracy one can consider the matter charge symmetric and neutral. Under these conditions, relations from § 8.7 and § 7.3 of Volume 1 apply.

What is the total number of different kinds of particles and antiparticles? Accelerator experiments demonstrate the existence of a great number of so-called resonances, i.e., elementary particles with extraordinarily short lifetimes of the order of 10^{-23} s. Of course, one cannot observe such particles by the classical methods of following their trajectories from the place of formation to the place of disintegration. Rather, the existence of these particles is manifested by, for example, the appearance of a maximum or "resonance" in the $(\pi^+ p)$ scattering cross section at a definite energy. One can describe this maximum by saying that the formation of a new particle, denoted by Δ^{++}, occurs according to the reaction $p + \pi^+ \rightarrow \Delta^{++}$. The charge of this new particle is twice as large as the charge of a proton. The Δ^{++} particle immediately disintegrates into $p + \pi^+$ so that one observes only the scattering of a pion by a proton. A detailed study permits one to determine not only the mass and charge, but also other properties such as the spin of the hypothetical particle "resonance." For the Δ^{++}, for example, the spin is 3/2.

Long-lived (and stable) particles form "families," as for example the proton and neutron or the three pions, π^+, π^0, and π^-. Resonances also form the same sort of families. Today the table of different resonances includes more than 300 members. This table continues to grow and it is possible that expansion of the table is limited only by the power of accelerators and the accuracy of experiments. (In Mikhail Romm's remarkable film, "Ten Days of One Year," a newspaper hanging on a wall makes this appeal: "Let us discover a new particle on the second block.")

Semiempirical theories exist that predict the existence of groups of related particles whose mass grows in proportion to the square root of their spin. In the theory of hot matter, there exists an idea—the theory of Hagedorn—according to which one can consider all these particles as independent components and can apply to them the formulae of statistical

mechanics. Fortunately one can understand the qualitative features of this idea without using the detailed theory set forth in Volume 1.

What happens is that in the formulae

$$\epsilon = \kappa\sigma T^4, \qquad S = \tfrac{4}{3}\kappa\sigma T^3, \tag{6.3.1}$$

where S is the entropy density and κ is the number of kinds of particles, it is necessary to consider κ as a function of temperature. The extreme case corresponds to $\kappa \to \infty$ for $T \to T_C$, where T_C is a certain critical temperature. A necessary condition for this is that the number of kinds of particles, multiplied by the statistical weight of each particle, grow exponentially with mass: If $\rho^*(m) \sim \exp(bm)$, then the limiting temperature is $T_C = c^2/kb$. The presence of power-law factors in front of the exponential does not change this basic result. The pressure is then less than that of a relativistic gas $P < \epsilon/3$, and in the extreme case when $\kappa \to \infty$, $P/\epsilon \to 0$. In the extreme case, consequently the cosmological expansion law turns out to correspond to that of the matter-dominated model:

$$a \sim t^{2/3}, \qquad \epsilon = \frac{c^2}{6\pi G t^2}. \tag{6.3.2}$$

Is such a picture realized? Indeed, it seems groundless to disregard the interactions in the situation considered. The very existence of a large number of resonances quite possibly indicates that we are concerned with the excited states of nucleons. Perhaps baryons are not elementary but composed of some "more elementary" particles. If this is so, then for high temperatures it is necessary to apply the principles of statistical physics to the latter particles. Suppose these "more elementary" particles are the hypothesized quarks. For a sufficiently high temperature, one can expect that $kT > m_q c^2$, where m_q is the mass of a quark. It is here assumed that quarks are much heavier than baryons; for example, $m_q = 30m_p$ or $m_q = 50m_p$. Then we return to the simple formula of the form (6.3.1) with the constant but small κ and, consequently, with $P = \epsilon/3$ also. Without entering into the essence of the matter, notice that Kirzhnits (1973) finds contradictions between the Hagedorn theory and modern field theory. The questions that arise in this area are quite complicated and require theoretical and experimental explanations. Existing physical theory does not answer these questions. What is more, it is not clear that these questions have an exact meaning. Let us explain the difficulties by means of a concrete example. Strongly interacting mesons possibly represent coupled baryon-antibaryon systems. As applied to pions, such a point of view was first expressed by Fermi and Yang in 1963. Later a classification scheme for all mesons (particles with zero baryon charge) was developed whereby the mesons are coupled states of quarks and antiquarks. But in dense, hot

matter, where large numbers of baryons and antibaryons are present, it is difficult to distinguish a baryon antibaryon pair from a meson.

Here it is necessary to emphasize that the difficulties and doubts relate to the possibility of a calculation of the number of individual particles, or to the use of the concepts of the statistical physics of noninteracting particles. The existence of concepts such as energy density and pressure, however, are not subject to doubt, and it is these quantities that enter into the right hand side of the equations of GTR. Moreover, there is no doubt that such thermodynamic quantities as temperature and entropy density also exist.

The laws of "charge" conservation, established for the pairwise interaction of particles, must also hold for dense matter.

In this sense the indeterminacy of the theory is often of a quantitative character. It is important to notice that the indeterminacy in the region of high temperatures, which exist at the beginning of the expansion of the Universe for $t < 10^{-6}$ s and $\rho > 10^{17}$ g cm^{-3}, is scarcely reflected subsequently—thanks to the rapid establishment of thermodynamic equilibrium. An exception would arise if, at a definite temperature, the homogeneous matter separates into two phases, one with an excess of nucleons and the other with an excess of antinucleons. Such a situation accords with the hypothesis of Omnès (see below). In this case the kinetics of the establishment of equilibrium depends on diffusion processes.

Hypotheses concerning the interaction of baryons can turn out to be important for cosmology. Considering the present level of knowledge, it is characteristic that two contrasting hypotheses appear in the literature. According to the first, there exists a repulsion of baryons from one another and an attraction between baryons and antibaryons. Thus the assumed theory is similar to electrostatics, but with these differences: (1) baryonic charge (baryon number) plays the role of electric charge and (2) the force has limited range: instead of the Coulomb potential $\sim r^{-1}$, the potential varies as $\sim r^{-1} \exp(-\mu r)$. Such a theory is obtained from the assumption that a neutral vector field exists that interacts with the baryonic charge; this differs from the electromagnetic field in that the quanta of this field—"vectons"—have a large rest mass.

For cold matter, this hypothesis (Zel'dovich 1961) leads to the equation of state

$$\epsilon = Mc^2 \left[n + \left(\frac{\hbar}{Mc} \right)^3 n^2 \right], \tag{6.3.3}$$

where n is the baryon density and M is the mass of a baryon. This is the so-called limiting, hard equation of state. Neither the comparatively weak, "ordinary" nuclear forces nor the Fermi energy of the nucleons is accounted for in this equation. However, if the assumption regarding repulsion is

true, then this neglect is unimportant in the limit $n \to \infty$. For cold matter, the repulsion can become substantial at a density

$$n \approx \left(\frac{Mc}{\hbar}\right)^3 \approx 10^{41}\ \mathrm{cm}^{-3}, \qquad \rho \approx 2M\left(\frac{Mc}{\hbar}\right)^3 \approx 4 \times 10^{17}\ \mathrm{g\ cm}^{-3}. \quad (6.3.4)$$

If hot matter contains an excess of baryons, there being $\mathcal{S} \approx 10^8$ pairs for each excess baryon, then the energy of repulsion is $\epsilon_{\mathrm{rep}} \approx (\hbar^3/Mc^2)n^2$. The thermal energy is $\epsilon_{\mathrm{th}} \approx N(\hbar c\, N^{1/3})$, where $N = \mathcal{S}n$, so that $\epsilon_{\mathrm{th}} \approx \hbar c\, \mathcal{S}^{4/3} n^{4/3}$. From this we find that value of n at which the repulsion becomes predominant:

$$n \approx \mathcal{S}^2\left(\frac{Mc}{\hbar}\right)^3 \approx 10^{57}\ \mathrm{cm}^{-3}, \qquad \rho \approx \mathcal{S}^4 M^4 c^3 \hbar^{-3} \sim 2 \times 10^{49}\ \mathrm{g\ cm}^{-3}. \quad (6.3.5)$$

For higher density, the equation of state $P = \epsilon$ leads to $a \sim t^{1/3}$. The deductions presented above are connected with an enormous extrapolation, and that is why they should in no way be considered reliable. On the other hand, notice that the characteristic "Planck" density $\rho_{\mathrm{Pl}} = c^5 G^{-2} \hbar^{-1} = 5 \times 10^{93}\ \mathrm{g\ cm}^{-3}$ is still much larger. As long as the density is less than ρ_{Pl}, the concepts of density and pressure can be used even if one is not able to calculate these quantities accurately.

A contrasting hypothesis concerning the interaction of baryons and antibaryons was recently advanced by Omnès (1969). On the basis of a theoretical analysis of the baryon-antibaryon interaction, Omnès maintains that antinucleons are repelled by nucleons in a definite energy range—and consequently in a definite temperature interval—of the order of several hundred MeV.

Following Omnès, let us construct the equation describing the equilibrium concentration of baryons x and antibaryons y. As usual, we will consider the charge symmetric system, so that the chemical potentials equal zero: $\mu_x = \mu_y = 0$. The energy of a baryon contains a term proportional to the antibaryon density,

$$E_x = [(Mc^2)^2 + p_x^2 c^2]^{1/2} + Ay. \quad (6.3.6)$$

Similarly, by symmetry,

$$E_y = [(Mc^2)^2 + p_y^2 c^2]^{1/2} + Ax. \quad (6.3.7)$$

From this we find that in equilibrium

$$x = B(T)e^{-Ay/kT}, \qquad y = B(T)e^{-Ax/kT}, \quad (6.3.8)$$

where $B(T)$ is the equilibrium concentration when account is not taken of the repulsion. This system of equations always has the symmetric solution $x = y = c(T)$, where $c(T)$ is the solution of the transcendental equation $c = B \exp[-Ac/kT]$. If, however, the interaction energy for one baryon is

greater than the thermal energy, i.e., if $b/kT > 1$ in the symmetric solution, then the system of two equations has the three solutions

$$x = y = c\,; \quad x_1 > c\,, \quad y_1 < c\,; \qquad x_2 = y_1 < c\,, \quad y_2 = x_1 > c\,. \tag{6.3.9}$$

Physically, this signifies that at high temperature, the repulsion between the nucleons and antinucleons leads to a separation of the charge symmetric matter into two phases. In one phase the nucleons predominate, and in the other the antinucleons predominate. The locally symmetric solution becomes unstable. Sufficiently hot matter ($T \gtrsim 300$ MeV is necessary according to the Omnès estimate) transforms into an emulsion, a mixture of two phases, the formation of large drops being thermodynamically favored. Charge symmetry is maintained only on the average over a volume containing many drops with an excess of baryons and many drops with an excess of antibaryons. With decreasing temperature, a single, homogeneous phase becomes thermodynamically favorable, but for it to arise the process of diffusion must smooth out the concentrations of baryons and antibaryons. Omnès thinks that the formation of two phases at a definite stage can be decisive for the charge-symmetric cosmology. It might lead to the separation of matter and antimatter on astronomical scales. We will turn to this side of the question in later chapters of the book.

Finally, there exists the assumption that the particles known to us—the proton, the neutron, and the resonances—are not elementary in the sense that they consist of some other elementary particles (Gell-Mann 1964; Zweig 1964). Several years ago the hypothesis of the quark was widely discussed. Of the three types of quark, all have baryonic charge $\frac{1}{3}$, while one has electric charge $\frac{2}{3}$ and the two others electric charge $-\frac{1}{3}$. By virtue of charge symmetry, this means that antiquarks also exist and have baryonic charge $-\frac{1}{3}$ and electric charge $-\frac{2}{3}$, $\frac{1}{3}$, and $\frac{1}{3}$.

In agreement with the quark hypothesis, each baryon consists of three quarks (in another variant [Zel'dovich and Sakharov 1965], each consists of four quarks and one antiquark). The charges are thus selected in such a way that the electric charge of a baryon can only have an integral value. In this scheme, a meson is a combination of a quark and an antiquark. The hypothesis was very successful because the various particles have an obvious explanation as distinct combinations of quarks. Each particle group is constructed from a distinct group of quarks with a particular spin orientation. The quark spin is $\frac{1}{2}$.

Searches undertaken for particles with fractional charge involved examinations of accelerator products, cosmic rays, and even small impurities in everyday matter. Regarding the searches for quarks in everyday matter, great significance was attached to predictions based on the hot model of the Universe (see § 7.4).

The search for quarks has not been crowned with success. At the same time, internal difficulties in the theory of quarks have been revealed. These difficulties apply specifically to the most economical model—the one with three kinds of quarks. In order for the model to agree with the systematics of the known particles, one must assume that quarks do not obey the Pauli principle and are not described by Fermi statistics, notwithstanding that they have spin $\frac{1}{2}$.

In a variant of the theory developed by Bogolyubov, Struminsky, and Tavkhelidze (1965), the existence of $3 \times 3 = 9$ kinds of quarks is assumed. In this variant the quarks, with spin $\frac{1}{2}$, are fermions. In addition, one can ascribe to them an integral electric charge. Therefore, experiments in which fractional charges are not found do not contradict such a variant of the theory. More recently, Dolgov *et al.* (1973) have constructed curious geometric models for the particles in such a theory. Recently proposals have appeared suggesting that hadrons consist of partons (from the Latin word "pars" meaning part). Of course, in any theory of the kind where ordinary hadrons disintegrate at a sufficiently high temperature, it is necessary to examine the properties of a gas consisting of quarks and antiquarks or partons and antipartons.

The particular aspects of the theory of Hagedorn and that of Omnès will be discussed in more detail below. Here, within the realm of a general review, one can only list the great diversity of variants and thus also the large indeterminacy in our knowledge concerning the state of matter for hadron and still higher temperatures.

One natural question immediately arises: if the indeterminacy in the past is so great, can theory predict the behavior of the subsequent stages of evolution?

Here conservation laws aid us. These are the exact conservation laws for energy and baryonic charge and the approximate conservation law for entropy during a slow, adiabatic process. We solve the problem of the evolution for the known values of the conserved quantities. These values are obtained from observations and relate to the current epoch; unfortunately, they have not been calculated theoretically. Since one makes use of experimental constants, however, the basic properties of the present-day situation and of that of the recent past may be accurately reproduced.

Deviations from the equilibrium state continue to be a problem for the theory. Specifically, the question of the number of baryons surviving to the present day is one of the most interesting. Analogously, one can pose the same question for the number of quarks and antiquarks or partons and antipartons. It is evident that the equilibrium number for a temperature of 3 K is inconceivably small. The answer in fact depends on the kinetics of the processes which govern the disappearance of the aforementioned particles. At the same time it is only necessary to consider particles that are stable in vacuum.

ow, an antiproton disappears only by annihilating with a proton. If quarks exist, then one kind of quark is stable and disappears only upon colliding with another quark of the same kind or with its antiquark. Thus we are dealing with processes requiring collisions between two particles. These processes occur slowly for a low density of particles.

Under these conditions the conclusions of the theory become firm, depending little on the indeterminacy of the theory for the hadron state. Decisive, conclusive accord occurs for much lower temperatures. Applicable to all of the variants are the assumptions about the spatial homogeneity of the Universe and the assignment of the density of baryon charge. In other words, a study of these variants constitutes an examination of the charge asymmetric Universe.

Let us now enumerate the conclusions resulting from the calculations that will be presented later in § 7.3 and § 7.4.

1. The residual concentration of antinucleons in the charge-asymmetric theory is extraordinarily small. This is so as early as the end of the hadron era, when the temperature is $\sim$1 MeV. This result follows naturally because (*a*) we are dealing with the strong interaction, so the annihilation cross section is large, and (*b*) the excess of nucleons makes for an exponentially small number of antibaryons when their creation ceases. The situation is the same as it is today in our neighborhood. Antinucleons and positrons are few—less than the number that cosmic rays produce after a cosmologically long time, because they annihilate with the excess nucleons present elsewhere.

2. The residual concentration of quarks turns out to be comparatively large. With respect to photons, it is of order $[GM^2/\hbar c)]^{1/2} \approx 10^{-18}$. With respect to ordinary nucleons, the concentration of quarks must be about 10^{-9}. This quantity is small in comparison to unity, but sufficiently large to inspire experimentalists. Of course, the calculation has meaning only if quarks exist. But now one can also say that if quarks are not observed in concentrations 10^{-15} to 10^{-20}, as experimentalists maintain, then quarks do not exist.

3. In an attempt to examine the spatially homogenous, charge-symmetric Universe, we find a problem concerning nucleons that is similar to the one we found concerning quarks. A similar answer obtains, namely, that the concentration is of order 10^{-18} nucleons per photon. Such a result contradicts the observations by roughly 10 orders of magnitude, which is why it is necessary to consider the charge-asymmetric, homogeneous Universe. This is the model to which § 6.1 and § 6.2 relate.

In this regard the theory of Omnès stands by itself. He proceeds from a charge-symmetric theory in which the primordial homogeneity of the Universe is spontaneously destroyed on a microscopic scale. The strong interaction leads to a separation of the homogeneous matter into drops of matter and drops of antimatter with a size of order 10^{-3} cm at the time

$t = 10^{-6}$ s. As the temperature decreases, the tendency toward separation of the matter and antimatter disappears. For low temperatures, annihilation occurs as usual. However, the spatial separation of the matter and antimatter substantially slows the annihilation. The rate of annihilation now is limited by the transfer of nucleons and antinucleons to the boundary of the separate regions. Only at this boundary does annihilation occur.

Concerning the subsequent development, there are at the present time enormous qualitative disagreements. Omnès supposes that after the annihilation there is a surplus of matter in regions with a size of the order of that of galaxies. He also assumes there remain about 10^{-9} nucleons per photon. Meanwhile, in other regions of the same size, there are as many antinucleons. In this way, Omnès thinks, the charge-symmetric theory leads to a picture of the present-day Universe consisting of galaxies and antigalaxies. His calculations give two characteristic quantities: the ratio of the average density of nucleons (antinucleons) to the photon density, and the characteristic size of a region filled with matter or with antimatter.

In our opinion Omnès's assumption about the separation of baryons and antibaryons during the hadron stage leads to other, more trivial conclusions. We note simply that the separation occurs by means of the mixing of nucleons and antinucleons on small scales.

A consistent calculation of the separation and the following annihilation leads to a much smaller residual concentration of nucleons and antinucleons. Thus the agreement between the theory and observations disappears at the most important point—in relation to the present-day density of nucleons. Omnès's theory also predicts that the annihilation is prolonged, continuing for the duration of the RD stage, and even later. But the expected consequences of a prolonged annihilation period are not observed. On the whole we believe that the charge-symmetric Universe, even with account of the phase separation, does not agree with observations and must therefore be rejected.

Let us thus turn to the concept of the charge-asymmetric Universe: a Universe with an excess of baryons for the entire duration of the evolution. In the hadron era, the excess of baryons is relatively small in view of the large number of pairs. Therefore, the picture describing the separation into a baryon and an antibaryon phase remains without change in the hadron era if the first hypothesis of Omnès is correct. For $T = 300$ MeV, charge asymmetry manifests itself only in the small predominance the volume of the nucleon phase has over that of the antinucleon phase. However, the picture of the annihilation changes substantially. According to our estimates, the annihilation of hadrons ends at a temperature higher than 1 MeV. Under these conditions, there is an abundance of electrons and positrons. In this case the RR spectrum takes the equilibrium form. Finally, the present-day density of baryons will remain equal to the initially specified baryonic charge density.

It would, of course, be most interesting to attempt to find some trace of the suggested hadron era phase separation in the present Universe.

6.4 Hagedorn's Theory

A short outline of Hagedorn's theory appears in Volume 1. (See also Hagedorn 1965, 1969, 1973.) For coherence in the account, we recall that noninteracting particles (an ideal gas) whose rest mass is m and whose statistical weight is g give a contribution to the energy density

$$\epsilon = g\psi(T, m) = \frac{g}{(2\pi\hbar)^3} \int \frac{E(p, m)d^3p}{\exp(E/T) \pm 1}, \qquad (6.4.1)$$
$$E = [m^2c^4 + c^2p^2]^{1/2}.$$

Here, the temperature is expressed in energy units and Boltzmann's constant $k = 1$. The chemical potential is taken equal to zero. This is always valid for neutral particles, but is also valid for charged particles in the symmetric case when equality in the number of particles with positive and negative charges exists. The sign "+" or "−" in the denominator corresponds to the case of fermions or bosons, respectively.

In the presence of many kinds of independent particles, it is necessary to take the sum of similar expressions:

$$\epsilon = \sum_i g_i\psi(T, m_i). \qquad (6.4.2)$$

If the number of constituents is finite and there exists some greatest mass $m_{\max}$, then the behavior of $\epsilon(T)$ for large T is simple. For $T > m_{\max}c^2$,

$$\epsilon = T^4\Sigma\alpha_i g_i,$$

where $\alpha_b = \pi^2/(30\hbar^3c^3)$ for bosons and $\alpha_f = 7\pi^2/(240\,\hbar^3c^3)$ for fermions.

Hagedorn examines an infinite sequence of particles. If the masses of the infinite number of particles are bounded, $m_i < M$ for $i \to \infty$, then for any finite temperature the theory predicts an infinite energy density: $\epsilon = \Sigma g_i\psi(T, m_i) > \Sigma g_i\psi(T, M) = \psi(T, M)\Sigma g_i = \infty$. Evidently such a conclusion contradicts experiment. In the theory of Hagedorn, therefore, one examines a sequence of particles with no limit on the mass.

Let us introduce the average number density of particles for asymptotically large m, so that the number of kinds of particles with mass in the interval between m and $m + dm$ is $dN = n(m)dm$. The product of the density and the average statistical weight, $f(m) = g(m)n(m)$, enters into formula (6.4.2). Replacing the sum by an integral, we obtain

$$\epsilon = \int \psi(T, m)\, f(m)dm. \qquad (6.4.3)$$

Here, a whole group of delicate mathematical problems arises. Their delicate character played a noticeable role in the large number of theoretical works which followed Hagedorn's proposal.

We shall find the behavior of the function $f(m)$ which leads to the existence of an upper limit on the temperature T, i.e., leading to the limit $\epsilon \to \infty$ as $T \to T_c$. In this situation, it is evident that the main contribution to the integral comes from the region $m > T$, where the asymptotic behavior is

$$\psi(T, m) = aT^{3/2}m^{5/2} \exp(-mc^2/T) . \qquad (6.4.1)$$

Hence, what must the function $f(m)$ be in order that $\epsilon = aT^{3/2} \int m^{5/2} \exp(-mc^2/T) f(m) dm \to \infty$ as $T \to T_c$? Clearly, the main condition is $f(m) \approx \exp(mc^2/T_c)$, or, more precisely, $f(m) = \phi(m) \exp(mc^2/T_c)$, where $\phi(m)$ is a slowly varying function. Then, $\exp[mc^2(T_c^{-1} - T^{-1})]$ appears in the integral. For $T < T_c$, the exponential decreases as m increases and the integral converges. For $T > T_c$, the exponential increases and the integral diverges. The form of the characteristic function $f(m)$ for which there exists a critical, maximum temperature T_c is thus determined. Moreover, one can define the behavior of $\epsilon(T)$ for $T \to T_c$ with respect to the slowly varying function $\phi(m)$. However, we will not occupy ourselves with this mathematical indulgence and instead turn to the physical basis of the theory.

Are the two assumptions upon which the theory is based, that of noninteracting particles and that of the lack of a bound on the number of types of particles, compatible?

Modern particle theory does not exclude such a situation. From experiment we know about the existence of the electron and the muon. Both of these particles possess electric charge and participate in "weak" interactions such as β-decay. Since the coupling constants are small—$e^2/(\hbar c) = 1/137 \ll 1$, $gm_\mu^2 c/\hbar^3 \approx 10^{-7} \ll 1$—the electron and muon can be considered as noninteracting particles at sufficiently high temperatures.

No modern theory has yielded a method for calculating the mass ratio $m_\mu/m_e = 207$. The existence of the muon itself is taken as an empirically established fact in the theory. Therefore, one cannot exclude the possibility that the electron-muon family has a continuation in a region of still higher mass (perhaps there exist superheavy muons and so forth), a region difficult to observe by experiment but nevertheless entering into statistical sums. However, it appears most unlikely that the number of members of such a family would be exponentially large.

The originator of the theory—Hagedorn—was inspired by the experimental data regarding the resonances of strongly interacting particles. Here one observes a regularity: the resonances with given charges (electric charge, baryonic charge, strangeness) lie on a straight line with respect to the coordinates S, the spin of the resonance, and m^2, the square of the mass of the resonance. On such a trajectory, the level density is $n(m) = \text{const.} \cdot m$. Meanwhile, the average statistical weight is $2S + 1 \sim m^2$, so that finally, $f(m) \sim m^3$. For a given trajectory, therefore, the function $f(m)$ grows, but more slowly than an exponential. With an increase in mass, new

particles can begin that correspond to new families of particles with nonzero isotopic spin.

In the case of hadrons, the possibility for exponential growth of $f(m)$ is greater than in the case of leptons (electrons and muons). Here, however, strictly *a priori*, one cannot consider the particles noninteracting in a hot dense plasma. Moreover, an account of this interaction in statistical terms is quite difficult to formulate. We will also encounter this question later during a more detailed discussion of the Omnès theory.

Suppose we examine a system in which pions occur along with protons and neutrons. For a calculation of the thermodynamic quantities relevant to such a system, it is necessary to take account of the phase space for scattering of a pion from a proton or a neutron. But a resonance—for example, the so-called Δ_{33} resonance—just manifests itself in the specific behavior of the cross section and of the phase space for scattering. By summing separately (*a*) pions interacting with nucleons and (*b*) resonances, one would err by accounting for the same physical phenomenon twice.

Hagedorn considers that the existence of a limiting temperature is confirmed by experiment for the following reasons. For any collisions of energetic elementary particles (in cosmic rays or at the most powerful accelerators) the formation of a "fireball" is assumed. The energy is isolated in a small volume, and a plasma in thermodynamic equilibrium arises. Hagedorn notices that the momentum of the particles flying away from the fireball corresponds to a temperature of 160 MeV, a figure practically independent of the number of particles created in any particular collision, i.e., independent of the total energy in the collision.

This argument is, however, not convincing because the observable particles probably do not immediately fly away from the volume in which the energy is originally released. In studies by Landau, Pomeranchuk, and their collaborators, other stages in collision of particles are also examined. In the first stage, a blob of dense plasma forms. Then this plasma expands according to the laws of hydrodynamics, with the total entropy conserved and the temperature decreasing. Finally, only in the last stage, when the density of the plasma becomes sufficiently small, do the observable particles fly away. Consequently Hagedorn's figure of 160 MeV is the temperature at which the plasma becomes transparent (transparency for hadrons also occurs at a size of the order of 10^{-13} cm) and the escape of the particles occurs.

The modern theory of fields (the theory of elementary of particles) gives an important argument against the theory of Hagedorn. Consider the polarization of the vacuum due to the electromagnetic field and the nonlinearity of electrodynamics thereby induced that leads, in particular, to the scattering of light by light. In the language of Feynman diagrams, the question concerns the processes corresponding to diagrams *a* and *b* in

Figure 29. There, the incident and exiting wavy lines correspond to photons, and the internal, closed line corresponds to some kind of charged particles. The process in diagram *a* leads to a renormalization of the charge. The observable elementary charge e is connected with the "bare" charge e_0 by an expression of the form

$$e^2 = \frac{e_0^2}{1 + \gamma(e_0^2/\hbar c) \ln (p_m/mc)},$$

where $\gamma = (3\pi)^{-1}$ is a dimensionless factor, p_m is the maximum momentum to which the theory is applicable, and m is the mass of the particle. The renormalization of charge is not directly observable, so that e_0 is not known. However, diagram a in Figure 29 also causes a small, observable deviation

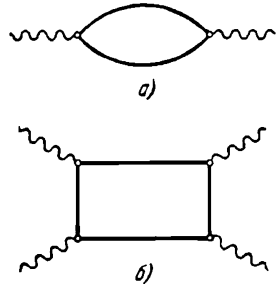


FIG. 29.—Feynman diagrams for the vacuum polarization due to an electromagnetic field.

in Coulomb's law. Diagram *b* leads to a cross section for the scattering of light by light. At the present time, the effects depending on the contribution of electrons and positrons have been experimentally verified. In reality, though, sums over all charged particles enter in the formulae corresponding to diagrams *a* and *b*. Morever, all terms of the sums have the same sign. From the evident facts that the observed charge is finite, that the cross section for the scattering of light by light is finite, and that, hence, the sums converge, one concludes that the number of types of charged particles is finite. These results contradict just those assumptions Hagedorn finds necessary!

Strictly speaking, considerations connected with the polarization of the vacuum relate to charged particles. The thermodynamic properties of the plasma depend equally on charged and neutral particles. In the case of hadrons, however, one cannot imagine that the number of types of neutral particles is infinite if the number of types of charged particles is finite. Thus, considerations connected with the polarization of the vacuum offer a telling argument against the theory of Hagedorn.

6.5 The Equilibrium Concentration of Nucleons and Antinucleons in the Charge Asymmetric Universe

In the region of high temperature, there exists a significant indeterminacy in the detailed physics due to the presence of large concentrations of strongly interacting particles—hadrons. An account of their interaction (which is strong, as the name of these particles itself indicates) is at the present time nonexistent. With the lowering of the temperature, the concentrations of hadrons decrease and the conclusions of theory become more definite and clear. When the density of baryon-antibaryon pairs becomes comparable to the density of excess baryons, it becomes immaterial whether we are dealing with a charge-symmetric or charge-asymmetric model of the Universe.

Thus, let us write the formulae for $kT < mc^2$, when the physics is reasonably clear. Dropping numerical factors, we have that

$$n_\gamma \sim \left(\frac{kT}{\hbar c}\right)^3 , \tag{6.5.1}$$

$$n\bar{n} \sim \left(\frac{kT}{\hbar c}\right)^3 \left(\frac{mc}{\hbar c}\right)^3 \exp\left(-2mc^2/kT\right) . \tag{6.5.2}$$

It is convenient to examine the dimensionless quantities $b \equiv n/n_\gamma$, $\bar{b} \equiv \bar{n}/n_\gamma$, $\beta = b - \bar{b}$, $\theta = kT/mc^2 \approx T_{13}\,(\mathrm{K})$. Recall that $T_{13} \equiv T \times 10^{-13}$. In a charge-symmetric Universe $\beta \equiv 0$. In a charge-asymmetric Universe, $\beta = \text{const.} > 0$, and the constancy of β corresponds to the conservation of baryons even during the expansion. The baryon number density $n - \bar{n}$ falls in the same way as does the photon number density. In order of magnitude, $\beta \approx 10^{-8}$. In the new variables, we have

$$b\bar{b} = \theta^{-3}e^{-2/\theta} , \qquad \beta = b - \bar{b} , \tag{6.5.3}$$

from which it is easy to find that

$$b = \frac{\beta}{2} + \left[\frac{\beta^2}{4} + \theta^{-3}e^{-2/\theta}\right]^{1/2} , \qquad \bar{b} = -\frac{\beta}{2} + \left[\frac{\beta^2}{4} + \theta^{-3}e^{-2/\theta}\right]^{1/2} . \tag{6.5.4}$$

There are three solutions to (6.5.4) which are of interest. When $\beta = 0$, it is evident that

$$b = \bar{b} = \theta^{-3/2}e^{-1/\theta} . \tag{6.5.5}$$

For nonzero β there are two solutions. For large β ($\beta \gg 2\theta^{-3/2}\, e^{-1/\theta}$, we have

$$b \approx \beta , \qquad \bar{b} \approx \frac{\theta^{-3}e^{-2/\theta}}{\beta} \ll b . \tag{6.5.6}$$

Finally, for small β ($\beta \ll 2\theta^{-3/2}\, e^{-1/\theta}$, we obtain

$$b \approx \bar{b} \approx \theta^{-3/2}e^{-1/\theta} \gg \beta . \tag{6.5.7}$$

During the expansion the temperature, and hence θ, decrease; ultimately solution (6.5.6) will apply. For $\beta \approx 10^{-8}$ this occurs for $\theta \lesssim 0.04$. In the temperature region considered here, $T \lesssim 4 \times 10^{11}$ K, the equilibrium situation can be calculated with satisfactory accuracy since the concentration of strongly interacting particles is not large. On the other hand, the kinetics of the annihilation and creation of pairs is still sufficiently rapid so that equilibrium actually exists. Only for a lower temperature does the rate of annihilation turn out to be insufficient; then the concentration of antibaryons exceeds that given by equilibrium. However, we will examine this kinetic problem separately (see § 7.3).

THE KINETICS OF ELEMENTARY-PARTICLE PROCESSES

1 Neutrinos in the Theory of the Hot Universe

In the earliest stages of expansion of the hot Universe, neutrinos exist in thermodynamic equilibrium with other particles. As is noted in § 6.2, equilibrium requires that the rate of reactions establishing equilibrium be fast compared to the expansion rate. We will now examine this condition as it relates to neutrinos.

The creation of neutrinos ν and antineutrinos $\bar{\nu}$ proceeds mainly according to the reaction $e^- + e^+ \rightarrow \nu + \bar{\nu}$. The cross section appropriate for relativistic electrons and positrons is given by

$$\sigma_\nu \approx \frac{g^2 E^2}{\hbar^4 c^4}, \tag{7.1.1}$$

where $g \approx 10^{-49}$ ergs cm^3 is the weak interaction constant. Let us substitute in place of E the order-of-magnitude estimate kT. Then, using the relation (6.2.3) for $T = T(t)$, we obtain the dependence of the time τ, required for the establishment of equilibrium on the cosmological time t:

$$\tau \equiv \frac{1}{\sigma_\nu n c} \approx \frac{G^{5/4}\hbar^{13/4}}{g^2 c^{1/4}} t^{5/2}. \tag{7.1.2}$$

When τ exceeds t, the neutrinos become free and their interaction with other particles and with one another ceases. Equating τ and t, we find $t \approx 0.1$ s. This estimate is made with expressions from which numerical factors have been omitted. It is presented to show how the result depends on the various constants such as G and g.

For a more precise estimate, one can write the exact expression. We make use of the expression, calculated by Chiu, for the rate of energy transfer from electrons and positrons to neutrinos and antineutrinos. We have

$$W = 4.6 \times 10^{-66} [T(\mathrm{K})]^9 \text{ ergs s}^{-1} \text{ cm}^{-3}, \quad T > 3 \times 10^9 \text{ K}.$$

Now compare this to the equilibrium energy density for ν and $\bar{\nu}$,

$$\epsilon_\nu = 6.8 \times 10^{-15} [T(\mathrm{K})]^4 \text{ ergs cm}^{-3},$$

to obtain the time for the establishment of the equilibrium.

$$\tau(\mathrm{s}) = \frac{\epsilon}{W} = \frac{1.5 \times 10^{51}}{[T(\mathrm{K})]^5} = \frac{7}{[T(\mathrm{MeV})]^5}. \tag{7.1.3}$$

We then use the connection between T and t: $t = 1.7\kappa^{1/2}[T(\mathrm{MeV})]^{-2}$ and we find (for $\kappa = 4.5$)

$$T \approx 2\ \mathrm{MeV} \approx 4m_e c^2\,, \qquad \tau \approx 0.2\ \mathrm{s}\,. \tag{7.1.4}$$

We shall denote the "decoupling" time by τ_ν. As we shall see later in this section, in order to deduce the present-day density of neutrinos it is essential that the equilibrium between $\nu\bar{\nu}$ and e^+e^- be maintained only in that period when $T \gtrsim m_e c^2$.

Now consider some instant of time t_1 which lies between 10^{-4} s and τ_ν:

$$10^{-4}\ \mathrm{s} < t_1 < 10^{-1}\ \mathrm{s}\,. \tag{7.1.5}$$

We shall denote the temperature at time t_1 by T_1. At this instant, there exists complete thermodynamic equilibrium among all of the types of leptons, including all kinds of neutrinos and antineutrinos. From the condition of thermodynamic equilibrium, one can write the equilibrium energy densities for each species of particle. Since it is essential that $m_\mu c^2 > kT_1 > m_e c^2$, the density of electron-positron pairs is roughly the same as the density of photons, differing only by a statistical factor. The density of muons is meanwhile insignificant. One can find the appropriate exact formulae in Landau and Lifshitz (1964). Let us write these relations for the equilibrium energy densities of the various particles at the low-energy end of the interval:

$$\epsilon_\gamma : \epsilon_{e^\pm} : \epsilon_{\nu,\bar{\nu}} : \epsilon_{\mu^\pm} = 1 \times \epsilon_1 : \tfrac{7}{4}\epsilon_1 : \tfrac{7}{8}\epsilon_1 : 10^{-3}\epsilon_1\,. \tag{7.1.6}$$

Here, $\epsilon_1 = \sigma T_1^4$. The number density of the other remaining particles is negligibly small. We will discuss the particular question of gravitons below. We denote the baryon density at this instant by ρ_1.

At earlier times all types of particles are found in equilibrium. For higher temperatures, strongly interacting mesons and baryon-antibaryon pairs appear. Equilibrium among the strongly interacting particles, photons, and the charged leptons is established almost instantaneously. Consequently, there is total equilibrium at temperatures higher than T_1. At lower temperatures, we encounter the decoupling time τ_ν.

Remarkably, the energy of the neutrinos after decoupling differs little from their equilibrium value. We now prove this.

Suppose that the Universe contains a mixture of noninteracting neutrinos and photons. The equation of state for the photons is

$$P_\gamma = \frac{\epsilon_\gamma}{3} = \rho_\gamma \frac{c^2}{3}\,. \tag{7.1.7}$$

Consider a volume V, expanding with the Universe, so that $V \sim a^3$. The total energy of the photons in this volume varies as

$$dE_\gamma = d(\epsilon_\gamma V) = -P_\gamma dV = \frac{-\epsilon_\gamma}{3}\,dV\,, \tag{7.1.8}$$

from which it follows that

$$\epsilon_\gamma \sim V^{-4/3} \sim a^{-4}, \tag{7.1.9}$$

a result well known from thermodynamics.

For neutrinos, a question arises. Do the concepts of energy within a given volume and of pressure apply if the neutrinos propagate freely, passing through the volume and beyond without collisions? Consider a particular neutrino. As for any particle, its momentum, measured in a comoving system, decreases inversely with a (see § 3.1):

$$p_\nu = \frac{E_\nu}{c} \sim a^{-1}. \tag{7.1.10}$$

The total number of neutrinos in a comoving volume is conserved inasmuch as they do not interact. Consequently, their density decreases as

$$n_\nu \sim a^{-3}. \tag{7.1.11}$$

From this it follows that the energy density of the neutrinos (for brevity we refer here, and in the following, to the energy of neutrinos, though we mean the sum of the energy of neutrinos and antineutrinos) decreases as a^{-4}:

$$\epsilon_\nu = n_\nu E_\nu \sim a^{-4}. \tag{7.1.12}$$

Thus, the law describing the decrease in the energy density of relativistic particles is the same for interacting particles as for noninteracting particles.

A study of the redshift of the particles is thus equivalent to following the thermodynamics of an adiabatic expansion. In § 3.11 the kinetic equation for photons in an expanding Universe was considered in detail. There it was shown that an equilibrium distribution remains in equilibrium; only a parameter of this distribution, the temperature, changes. The same is true here for neutrinos.

Let us account for the continued interaction between the photons and the electrons and positrons after the decoupling of the neutrinos. It is not difficult to find the specific entropy for the neutrinos and for the combination of photons and electron-positron pairs at the time when $T = T_1$:

$$\begin{aligned} S_\gamma + S_{e^\pm} &= \frac{4}{3}\left[\frac{\sigma T_1^3 + (7/4)\sigma T_1^3}{\rho_1}\right]; \\ S_{\nu_e} = S_{\nu_\mu} &= \frac{4}{3}\cdot\frac{7}{8}\cdot\frac{\sigma T_1^3}{\rho_1}. \end{aligned} \tag{7.1.13}$$

In the course of the expansion, the part of the entropy depending on the neutrinos and the part depending on the photons and the $e^\pm$ pairs are separately conserved. However, the photons and $e^\pm$ pairs are in a dynamic equilibrium and exchange energy and entropy. When $T \lesssim m_e c^2 \approx 5 \times 10^9$ K (i.e., when $t \gtrsim 10$ s), the electron-positron pairs annihilate into

photons. The entropy of the pairs passes to the photons. After this, the expansion leads (after 10^{10} years!) to the present-day situation. Denoting the present-day situation by the index zero, we obtain

$$S_{\gamma,0} = S_{\gamma,1} + S_{e^{\pm},1} \quad \text{or} \quad \frac{4}{3}\cdot\frac{\sigma T_{\gamma,0}^3}{\rho_0} = \frac{4}{3}\cdot\frac{11}{4}\cdot\frac{\sigma T_1^3}{\rho_1}\,; \tag{7.1.14}$$

$$S_{\nu_e,0} = S_{\nu_e,1} \quad \text{or} \quad \frac{4}{3}\frac{7}{8}\frac{\sigma T_{\nu,0}^3}{\rho_0} = \frac{4}{3}\cdot\frac{7}{8}\cdot\frac{\sigma T_1^3}{\rho_1}\,.$$

Here $T_{\gamma,0}$ is the present-day photon temperature of 2.7 K, and ρ_0 is presumably $\sim 10^{-29}$ g cm^{-3}. Independent of the numerical value of ρ_0 presented above, it follows that the present temperatures and energy densities of both kinds of neutrino are

$$T_{\nu,0} = \left(\frac{4}{11}\right)^{1/3} T_{\gamma,0} \approx (0.7) \times 2.7\ \text{K} \approx 2\ \text{K}\,, \tag{7.1.15}$$

$$\epsilon_{\nu_e,0} = \epsilon_{\nu_\mu,0} = \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} \epsilon_{\gamma,0} = (0.35)\epsilon_{\gamma,0}\,.$$

This result was obtained by Peebles (1966*a*, *b*).

Thus, the theory of the hot Universe predicts that there exists today an equilibrium, relic, neutrino background with a temperature $T_\nu \approx 2$ K.

An experimental observation of an equilibrium spectrum of neutrinos would be of enormous interest. If the relation of $T_{\nu,0}$ to $T_{\gamma,0}$ were confirmed, then this would constitute proof of the correctness of our ideas about the earliest stages of expansion when $T \gtrsim 10^{10}$ K, $\rho \gtrsim 10^5$ g cm^{-3}. According to a remark by A. D. Sakharov, the experiment could be based on the change in the form of the spectrum of β-decay electrons near the maximum caused by the presence of low-energy neutrinos. However, as B. M. Pontecorvo has noted, a small rest mass for the neutrino would have an analogous effect. Experiment gives the limit $m_{\nu_e}c^2 < 100$ eV. However, the cosmic background consists of neutrinos with an energy of the order of 5×10^{-4} eV. Hence, in order to observe these neutrinos it is necessary to improve the accuracy of the experiment by 10^6.

A curious situation arises for muon neutrinos. The methods of elementary-particle physics permit only an imprecise determination of their rest mass; the limit is $m_{\nu_\mu} < 2\ \text{MeV}/c^2$. However, from cosmological considerations, i.e., knowing that a number density of muon neutrinos of the order of 200 cm^{-3} must be present today, and knowing the maximum permissible density of all forms of matter in the Universe, one can obtain a significantly better estimate, $m_{\nu_\mu} < 500\ \text{eV}/c^2$ (Gershtein and Zel'dovich 1966). Later this method was used to obtain the more accurate estimate that $m_{\nu_\mu} \lesssim 100$ eV.

Research on the photon spectrum permits verification of the homoge-

isotropic, hot model back to redshifts $z \approx 10^6$. This corresponds to about 1 yr after the singularity (see chap. 8).

Conceivable, research on the spectrum of neutrinos and antineutrinos will give an indication of deviations from the classical model at a time of a few tenths of a second after the singularity, at a redshift $z \approx 10^{10}$. (Also in Part IV, which is dedicated to the study of anisotropic solutions, the changes in the density and spectrum of the neutrinos in the corresponding cosmological models will be examined.

Above we saw that the comparatively early decoupling of the neutrinos from the other particles decreases the present-day neutrino energy with respect to the present-day energy of the RR photons. This happens because, in early stages, the energy is distributed among a large number of kinds of particles. Such reasoning imparts great stability to the whole theoretical picture. If some other, yet unknown particles exist that interact with the known particles still more weakly than do the neutrinos, then they would decouple still earlier and consequently would contribute substantially less to the present total energy density than do the neutrinos. In particular, photons apparently play a small role for this reason. However, see the next section for a more detailed discussion of them.

Evidently, one can conceive of unknown, strongly interacting particles only if they have large masses—otherwise they would already have been discovered! Indeed, quarks constitute a particular case of such unknown, hypothetical particles. Such heavy, strongly interacting particles of mass m would change the relation between T and ϵ for $T > mc^2$. At the time when $T \approx mc^2$ is reached, however, they would become extinct, having no influence at all on the subsequent evolution of the Universe.

Particles capable of spontaneously disintegrating disappear exponentially with time as the temperature decreases and their formation ceases; observation of such relic particles is impossible.

Things are different in the case of particles that are stable in a vacuum and that disappear only upon interacting with other particles. Examples of these are antiprotons $\bar{p}$, positrons e^+, and, if quarks exist, one single type of them (see the end of the chapter for an estimate of their density).

An overall picture of the change in the relative densities of various particles in the expanding model is shown in Figure 30.

7.2 Cosmological Gravitational Waves

Gravitational waves differ from neutrinos in that their interaction with matter is much weaker. At the present time there is no doubt that gravitational waves exist as a consequence of GTR. This conclusion does not depend on the results of Weber's experiment or on similar work provoked by his articles (see the review in Vol. 1).

The quantization of gravitational waves is in principal a necessary part

of quantum theory. Thus the concept of the graviton—a quantum of a gravitational wave—has appeared, with the graviton carrying an energy $\hbar\omega$. The analogous concepts in Maxwell's theory are electromagnetic waves and photons. The analogy goes still further: gravitons as well as photons have zero rest mass, and both are transverse with two independent states of polarization. In Maxwell's equations, the long-range Coulomb field is not quantized; the same holds for the long-range components of the gravitational field. In thermodynamic equilibrium, the energy density of gravitational radiation equals the energy density of electromagnetic radiation. The equilibrium spectrum of gravitons obeys the Planck formula exactly.

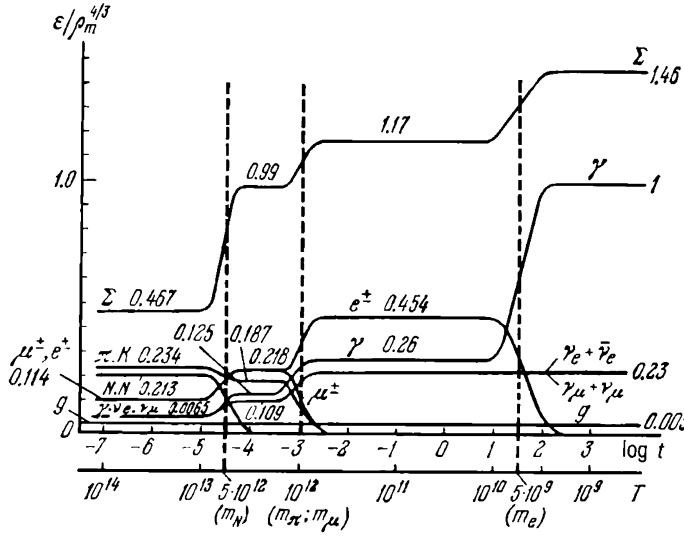


FIG. 30.—Evolution of $\epsilon/\rho_m^{4/3}$ for various kinds of particles in the hot, expanding model of the Universe. The quantity ϵ is the energy density, while $\rho_m = m_{\text{nuc}}\, n$, where m_{nuc} is the nucleon rest mass and n is the baryon number density (density of baryons minus density of antibaryons). Four forms of baryons and four forms of antibaryons respectively are encompassed by the symbols N and $\bar{N}$; $\epsilon_\gamma/\rho_m^{4/3}$ is taken equal to unity at the present time; and Σ signifies a summation over all particles.

So how does gravitational radiation differ from electromagnetic radiation? One difference is theoretical: gravitational radiation represents a change in the metric of spacetime. It is often asked whether we can consider this radiation *on a background* of a classical metric. To this question one gives an answer concerning the technique of separating the metric into slowly changing and rapidly changing quantities. With respect to a connection with the theory of the stress-energy pseudotensor, the notion goes back to Landau and Lifshitz's *Theory of Fields* (1962); it was also realized by Isaacson (1968*a*, *b*; see also Vol. 1). For quantization it is essential that the long-range gravitational field not be quantized. This is what allows one to

consider gravitons on a background of a classical, nonquantum, averaged metric.

The practical, quantitative, distinguishing feature of gravitational radiation is that its interaction with matter is exceptionally small, especially on the level of elementary particles. The emission of electromagnetic waves is proportional to the square of the charge: e^2. This means that the dimensionless factor characterizing the probability of emission is $\alpha = e^2/\hbar c = 1/137$. For example, let a charged particle be deflected upon colliding with another charged particle. The probability that this process is accompanied by the emission of one photon equals the probability of the nonradiative process multiplied by α. Other factors—v/c and r/λ—can also enter, but the factor $\alpha = 1/137$ is mandatory!

For the radiation of a *gravitational* wave (graviton), the factor $Gm^2/\hbar c$ plays an analogous role. This is not surprising, for one need only compare the expressions for the Coulomb energy of two charges e^2/r and the gravitational energy of two masses Gm^2/r to convince oneself that Gm^2 plays the role in gravitational theory that the square of the charge e^2 plays in electromagnetic theory. Now, the quantity $Gm^2/\hbar c$ equals $\sim 10^{-38}$ for protons and 10^{-44} for electrons. The small size of this quantity leads to the conclusion that the emission, and absorption, of gravitational waves is negligibly small. Thermal gravitational radiation is thus not observable in the laboratory. It is evident that a laboratory measurement of the thermal radiation inside a "blackbody" yields an equilibrium density of electromagnetic radiation. However, nothing similar to an equilibrium density of gravitational radiation ever occurs. We are always dealing with incomplete thermodynamic equilibrium.

In laboratory experiments, neither gravitons nor neutrinos nor antineutrinos are in thermal equilibrium. Upon passing to astrophysics, to large spatial scales and much higher densities and temperatures, the creation and absorption of weakly interacting particles become more important. In hot stars, the emission of neutrinos and antineutrinos becomes significant, but their density is still small with respect to their density in an equilibrium situation since the neutrinos can leave the star freely. In stars undergoing collapse (transforming into neutron stars or black holes), conditions are reached in the center such that the mean free path of neutrinos is so small that they leave by diffusion, experiencing many collisions. Under such conditions, thermodynamic equilibrium between the hot ordinary matter and the neutrinos can be established. But under these same conditions, the interaction of gravitational waves with matter is negligibly small, the density of gravitons is small compared to its equilibrium value, and one can still ignore the energy loss due to gravitons.

Let us now turn to a consideration of the epoch near the singularity. We will be studying the isotropically expanding, homogeneous cosmological model. We restrict attention to the hot model ($S \approx$ const.) containing a

finite number of kinds of particles, and we will assume that the rest masses of the particles are less than or of the order of that of a baryon. In other words, we are not taking account of Hagedorn's ideas.

With a finite number of kinds of particles, one can consider the matter to be ultrarelativistic at an early stage. An order-of-magnitude estimate of the density is

$$\rho = \frac{\epsilon}{c^2} \approx \frac{p}{c}\left(\frac{p}{\hbar}\right)^3 \approx (Gt^2)^{-1}, \tag{7.2.1}$$

so that the typical momentum is

$$p \approx \left(\frac{\hbar^3 c}{Gt^2}\right)^{1/4} \approx m_{\mathrm{Pl}} c \left(\frac{t_{\mathrm{Pl}}}{t}\right)^{1/2}. \tag{7.2.2}$$

In the above, $m_{\mathrm{Pl}} = (\hbar c/G)^{1/2} \approx 2 \times 10^{-5}$ g is the Planck mass and $t_{\mathrm{Pl}} = (G\hbar/c^5)^{1/2} \approx 5 \times 10^{-44}$ s is the Planck time. We may also write (7.2.2) in the form

$$\frac{p}{mc} \approx \left(\frac{Gm^2}{\hbar c}\right)^{-1/4} \left(\frac{\hbar/mc^2}{t}\right)^{1/2}. \tag{7.2.3}$$

From (7.2.3) it follows that nucleon-antinucleon pairs remain relativistic until $t \approx 10^{-5}$ s. For later times such pairs annihilate and there will be none in equilibrium. It is easy to convince oneself that even a negligibly small exchange of energy between ordinary matter and gravitons is proscribed even before 10^{-6} s. This is because the reaction time τ satisfies $\tau \gg t$, so the gravitons are completely "frozen." The temperature of such noninteracting gravitons decreases in the course of the expansion inversely proportional to the scale factor—just as for the neutrinos after decoupling.

The establishment of equilibrium between gravitons and the other particles could occur only at a much earlier stage—one in which an ultrarelativistic temperature prevails. At such a stage the energy and momentum of the particles are of the order of T and T/c, respectively, which are much greater than mc^2 and mc.

The energy and momentum of particles enter into the equations of GTR. Inasmuch as we are concerned with the collisions of particles, their rest mass is not important since it is small in comparison to their total energy. From this it follows that the probability of emission of a graviton must be proportional to the factor $GE^2/\hbar c^5$; this factor differs from the nonrelativistic factor $Gm^2/\hbar c$ presented above in that m is replaced by E/c^2.

The nongravitational interaction of particles in the ultrarelativistic limit is characterized by an interaction cross section of the order of $(\hbar/p)^2 = (\hbar c/E)^2$. Consequently, for the cross section for the creation of a single graviton, we obtain the estimate

$$\sigma_1 \sim \left(\frac{\hbar c}{E}\right)^2 \frac{GE^2}{\hbar c^5} \sim \frac{G\hbar}{c^3}. \tag{7.2.4}$$

If all that if the emission of a photon occurs at the same time, an additional factor $e^2/\hbar$ appears. We especially emphasize that in the ultrarelativistic case, the one graviton cross section depends neither on the rest mass nor on the energy of the interacting particles. This cross section is of essentially the square of the "Planck length":

$$\sigma_1 \approx l_{\mathrm{Pl}}^2 ; \qquad l_{\mathrm{Pl}} = \left[\frac{G\hbar}{c^3}\right]^{1/2} . \tag{7.2.5}$$

For the creation of two gravitons we find

$$\sigma_2 \sim \left(\frac{\hbar c}{E}\right)^2 \left(\frac{GE^2}{\hbar c^5}\right)^2 \sim l_{\mathrm{Pl}}^2 \left(\frac{GE^2}{\hbar c^5}\right) . \tag{7.2.6}$$

The situation is similar when $\nu\bar{\nu}$ pairs are created. For both possibilities, annihilation of particle-antiparticle pairs plays the main role in the ultrarelativistic case—not scattering.

Here we can draw on the results of the pioneering work of Vladimirov (1960), in which the annihilation cross sections for $e^- + e^+ \to \gamma + g$ and $e^- + e^+ \to 2g$ (γ represents a photon and g a graviton) are calculated in accordance with the rules of quantum electrodynamics. In the ultrarelativistic limit, the results take the form

$$\sigma_1 = \frac{1}{3}\frac{e^2}{\hbar c} l_{\mathrm{Pl}}^2 ; \qquad \sigma_2 = \frac{\pi}{3} l_{\mathrm{Pl}}^2 \frac{GE^2}{\hbar c^5} . \tag{7.2.7}$$

In the first expression the electric charge enters. The second expression is valid for any particles, charged or uncharged. The specific numerical factor applies to the annihilation of particles with spin $\frac{1}{2}$. When the energy of the particles is less than the Planck energy, i.e., when $GE^2/\hbar c^5 \ll 1$, the probability of a one-graviton process is larger. One can further suppose that for the process

$$p + \bar{p} \to \text{hadrons} + g , \tag{7.2.8}$$

the cross section in the ultrarelativistic limit is simply

$$\sigma_1 \approx \frac{G\hbar}{c^3} = l_{\mathrm{Pl}}^2 . \tag{7.2.9}$$

No "small" factor connected with the electromagnetic interaction enters.

With this knowledge about the probability for the creation of gravitons, let us try to determine the characteristic time t_G of graviton decoupling. The answer is evident in advance from dimensional considerations: in the ultrarelativistic region $t_G \approx t_{\mathrm{Pl}}$. Indeed, there is no other characteristic time which one can construct from G, $\hbar$, and c.

Particle rest masses do not enter into the result because we are considering an ultrarelativistic gas with $E \gg mc^2$. Further, the strong interaction is characterized by the dimensionless quantity $g^2/\hbar c$, which is of

order unity. In principle, the number of kinds of particles can enter into the result; however, we assume that this number, of order unity, is not larger than 10 or 20. Indeed, it is in this step that the renunciation of Hagedorn's hypothesis is important. Finally, there are dimensionless numbers such as $3/32\pi$ in the formula $\rho = 3/(32\pi G t^2)$, and similar numerical factors in the Planck radiation law, in the cross section expressions, and so forth. An account of these numbers and an attempt at a "more exact" calculation would have an illusory character.

Note also that we are actually concerned not with the interaction of two particles in a vacuum, but with their interaction in an extraordinarily dense plasma, for which a quantitative theory does not exist. For the present, therefore, we must be content with this rudimentary graviton theory.

The conclusion from dimensional considerations is thus that the establishment of equilibrium is not guaranteed since, by a rough calculation, $t_G \sim t_{Pl}$ and we agreed to consider t_{Pl} the lower limit on t for which the theory applies. This conclusion does not change even if the estimate of t_G changes by a large factor, because it is not known whether the boundary of the theory is at t_{Pl}, $5t_{Pl}$, or $0.2t_{Pl}$.

Practically simultaneously with the publication of the estimates presented above in *Relativistic Astrophysics* (Vol. 1), the work of Matzner (1968) appeared. In the latter, the opposite conclusion is drawn, namely, that equilibrium between gravitons and other particles prevails for a long time right up to $t = t_{Pl}(Gm^2/\hbar c)^{-2/3}$. The author comes to this conclusion by considering the scattering of the strongly interacting hadrons. Experiment shows that for large energies $E > mc^2$, the interaction cross section tends to a constant. This deduction agrees with modern asymptotic theory for particle scattering. But both the experiments and theory apply to interactions without the emission of a graviton. In order to obtain the cross section for the emission of a graviton, Matzner multiplies the cross section for the graviton-free interaction by $GE^2/\hbar c^5$. But it seems to us that at this point an important error is committed.

The point is that in the asymptotic limit $E/mc^2 \to \infty$ the constant cross section, not including graviton effects, is connected with a decrease in the scattering angle. The average transverse momentum transferred tends to a constant, but the angle—the ratio of the transverse momentum to the longitudinal—tends to zero. Meanwhile, the probability for the radiation of gravitons whose energy is of the order of the longitudinal energy depends on the momentum transferred. Hence, small-angle scattering gives a correspondingly small amplitude for gravitational-wave emission. Accounting for the decrease in the angle, I. Yu. Kobzarev arrives at the asymptotic form $\sigma = l_{Pl}^2$, i.e., he arrives at the conclusion used by us above. Thus, in contrast to Matzner, we conclude that the establishment of equilibrium between the gravitons and the other particles is not guaranteed. It is

[illegible] uncertain whether the density of gravitons is less than or greater [illegible] the equilibrium density.

[illegible] pulibrium at an early stage would lead to $\epsilon_G \approx 0.02\epsilon_\gamma$ at the present [illegible] we assume that the entropy of the hadrons is transferred to photons—[illegible] not to gravitons. This reasoning is analogous to that which Peebles [illegible] (1966b) used in analyzing neutrinos. The graviton density can be much [illegible] if the entropy of the hot Universe arises as a result of some kind of [illegible] tion process after t_{Pl}. However, the condition $\epsilon_G \ll 0.02\epsilon_\gamma$ does not lead [illegible] to change in the conclusions relating to nucleosynthesis, the growth of [illegible] perturbations, the formation of galaxies, and so on. In principle, another [illegible] scenario is possible: If the entropy depends on the spontaneous creation of [illegible] particles at an early, anisotropic stage of expansion, then one can expect [illegible] substantial creation of gravitons. Up to now we have spoken exclusively [illegible] about the creation of gravitons, but it is evident that in the ultrarelativistic [illegible] limit the cross sections for the forward and reverse processes are equal. [illegible] Hence, when the density of gravitons is of hyperequilibrium character, they [illegible] partially transfer their energy and entropy to the other particles, though a [illegible] hyperequilibrium (with respect to the temperature of the other particles) [illegible] [illegible] of gravitons remains. Indirectly, one can judge the possible limits on [illegible] the excess of gravitational radiation by its influence on nucleosynthesis. The [illegible] conclusion is that (see § 7.1) $\epsilon_G \lesssim 3\epsilon_\gamma$. This would result in $\rho_G \lesssim 1 \times 10^{-33}$ g [illegible] cm^{-3} and $\Omega_G \lesssim 10^{-4}h_0^{-2}$ today. Such a graviton background can in no way [illegible] influence the dynamics of the Universe.

During the RD period, a high graviton density would lead to a small [illegible] quantitative change in the expansion rate, in the time when the equality [illegible] $\rho_\nu = \rho_\gamma + \rho_G$ is satisfied, and in the time of hydrogen recombination (see [illegible] § 8.1). However, qualitative changes are not expected, so we will not take [illegible] account of this possibility in the subsequent exposition.

Gravitational waves of low frequency and large wavelength can be [illegible] examined classically without an account of quantum effects. This follows [illegible] because they interact coherently with matter in large volumes. A systematic [illegible] review of the various aspects of such gravitational radiation appears in a [illegible] separate chapter.

7.3 Antinucleons in the Hot Model

In connection with the hypothesis of a hot model it is most interesting from a methodological viewpoint to examine the kinetics of processes in one variant of the completely charge-symmetric model. In particular, this variant is the one in which there are equal numbers of nucleons and antinucleons (and also of electrons and positrons and so forth) and in which their densities are uniform. Hagedorn's ideas are not taken into account.

For a sufficiently high temperature, but one nevertheless less than Mc^2/k, the densities of the particles are fixed in accordance with the condition of

thermodynamic equilibrium, where account is taken of the nucleon's spin. The variable n_N is the sum of the densities of protons and neutrons, while $n_{\bar{N}}$ is that for antiprotons and antineutrons:

$$n_N = n_{\bar{N}} = \left(\frac{2}{\pi^3}\right)^{1/2} \frac{(MkT)^{3/2}}{\hbar^3} \exp\left(-\frac{Mc^2}{kT}\right). \tag{7.3.1}$$

Thermodynamic equilibrium exists because of the balance between the annihilation of colliding nucleons and antinucleons and pair creation. In the course of the expansion, the equilibrium density decreases, and for the present-day temperature of $\sim$3 K, this would be "astronomically" small, $\sim 10^{-10^{12}}$. However, the annihilation rate actually becomes negligible much earlier; and a "freeze-out" occurs with a definite, nonequilibrium number of nucleons and antinucleons remaining. Let us examine this process.

The annihilation cross section in the region of interest to us varies like $\sigma_0 c/v$, where v is the collision velocity and σ_0 is of the order of 10^{-26} cm^2. From this we obtain the equation for the change in density:

$$\frac{dn}{dt} = -\sigma_0 c n^2 + \psi(t) - \alpha n \,. \tag{7.3.2}$$

Here $\psi(t)$ is a function describing the pair creation rate. The last term, $-\alpha n$, represents the change in the density due to the expansion; in particular, $\alpha = 3H = 3/2t$. It is convenient to consider the photons (more precisely, the totality of light particles—$e^{\pm}$, γ) in parallel with the baryons. The density of the former obeys the law

$$\frac{dn_\gamma}{dt} = -\alpha n_\gamma \,, \tag{7.3.3}$$

which describes the change in density of conserved particles.

We now introduce the ratio $\eta = n/n_\gamma$. It is evident that the equation for η does not contain any dependence on the expansion. Hence, were there no annihilation or formation of new pairs, η would be constant. In fact,

$$\frac{d\eta}{dt} = -\sigma_0 c n_\gamma \eta^2 + \frac{\psi}{n_\gamma} \,. \tag{7.3.4}$$

We introduce η_{eq} such that

$$\eta_{eq} \approx \theta^{-3/2} e^{-1/\theta} \,; \quad \theta = kT/Mc^2 \,. \tag{7.3.5}$$

In equilibrium,

$$-\sigma_0 c n_\gamma \eta_{eq}^2 + \frac{\psi}{n_\gamma} \equiv 0 \,, \tag{7.3.6}$$

so that the equation for η can be written in the form

$$\frac{d\eta}{dt} = -\sigma_0 c n_\gamma (\eta^2 - \eta_{eq}^2) \,. \tag{7.3.7}$$

In equation (7.3.7), n_γ and η_{eq} are known functions of θ, while θ and η depend on t. We will separate the interval of time $0 < t < \infty$ into two parts. The first, $t < t_N$, is where approximate equilibrium is maintained, i.e., where $\eta - \eta_{eq} \ll \eta_{eq}$ holds; the second, $t > t_N$, is where $\eta \gg \eta_{eq}$.

To locate the characteristic time t_N, we set $\eta = \eta_{eq}$ on the left-hand side of the equation, calculate the difference $\eta - \eta_{eq}$, and demand that, in the limit, the condition $\eta - \eta_{eq} \approx \eta_{eq}$ be satisfied. Thus

$$\begin{aligned} \frac{d\eta_{eq}}{dt} &\approx -\sigma_0 c n_\gamma (\eta^2 - \eta_{eq}^2) ; \\ \eta - \eta_{eq} &\approx \frac{1}{n_\gamma \sigma_0 c (\eta + \eta_{eq})} \left| \frac{d\eta_{eq}}{dt} \right| . \end{aligned} \tag{7.3.8}$$

The condition $\eta - \eta_{eq} \approx \eta_{eq}$ gives $2n_\gamma \sigma_0 \eta_{eq} \approx \eta_{eq}^{-1} |d\eta_{eq}/dt| = |d \ln(\eta_{eq})/dt|$, or since

$$\begin{aligned} &\ln \eta_{eq} \approx \theta^{-1} \quad \text{and} \quad \theta \sim t^{-1/2} , \\ &\left| \frac{d \ln (\eta_{eq})}{dt} \right| = \frac{1}{2\theta t} \approx 2\sigma_0 c n_\gamma \eta_{eq} . \end{aligned} \tag{7.3.9}$$

Finally, the equation becomes

$$4\sigma_0 c \eta_{eq,N} \theta_N t_N n_{\gamma,N} \approx 1 . \tag{7.3.10}$$

substituting the numerical values of

$$t \approx 10^{-6} \theta^{-2} \text{ s} , \quad n_\gamma \approx 10^{41} \theta^3 \text{ cm}^{-3} , \quad \sigma_0 = 2 \times 10^{-26} \text{ cm}^2 ,$$

we then obtain the equation

$$e^{+1/\theta_N} \approx 2.4 \times 10^{20} \theta_N^{1/2} . \tag{7.3.11}$$

This equation is solved by successive approximations. We obtain as a final result

$$\begin{aligned} \theta_N = \frac{kT_N}{Mc^2} \approx \frac{1}{44} , \quad kT_N \approx 20 \text{ MeV} , \quad t_N \approx 2.5 \times 10^{-3} \text{ sec} , \\ \eta_N \approx \eta_{eq,N} \approx 2 \times 10^{-18} , \quad n_{\gamma,N} \approx 10^{36} \text{ cm}^{-3} , \\ n_N \approx n_{eq,N} \approx 2 \times 10^{18} \text{ cm}^{-3} . \end{aligned} \tag{7.3.12}$$

Thus, under the assumption of a charge-symmetric, homogeneous, and isotropic model, nucleons N and antinucleons $\bar{N}$ constitute a fraction 2×10^{-18} of the light particles at the time when thermodynamic equilibrium is destroyed. When $kT/Mc^2 \geq 1$, the density of nucleons is of the same order as the density of light particles. In the interval $1 \gtrsim kT/Mc^2 \gtrsim 1/44$, the density falls exponentially, and by the time $kT/Mc^2 \approx 1/44$, the density of nucleons is 2×10^{-18} of the density of the light particles.

Later, i.e., when $kT/Mc^2 < 1/44$ or $t > 2.5 \times 10^{-3}$ s, the equilibrium value of the nucleon density n_{eq} continues to decrease exponentially but equilibrium does not hold, so that the nucleon density falls more slowly. There-

fore, $n \gg n_{eq}$, and in the equation for dn/dt one can neglect the creation of new N-$\bar{N}$ pairs. Then, the only remaining term in the equation is that describing annihilation:

$$\frac{d\eta}{dt} \approx -\sigma_0 c n_\gamma \eta^2 . \tag{7.3.13}$$

Initial conditions are taken from the discussion just presented and are

$$t_N \approx 2.5 \times 10^{-3}\ \text{s}\ ; \quad \eta_N \approx 2 \times 10^{-18}\ ; \quad 4\sigma_0 c n_{\gamma N} \eta_N t_N \approx \theta_N^{-1} . \tag{7.3.14}$$

This equation is easily solved. Taking $\tau = t/t_N$ and $n_\gamma = n_{\gamma,N}\tau^{-3/2}$, we find

$$\frac{\tau^{3/2}}{\eta^2}\frac{d\eta}{d\tau} \approx -(4\eta_N\theta_N)^{-1} \approx (2 \times 10^{-19})^{-1} , \tag{7.3.15}$$

which may be integrated to yield

$$\begin{aligned} \frac{\eta}{\eta_N} &\approx [1 + 10^{19}\eta_N(1 - \tau^{-1/2})]^{-1} , \\ \frac{\eta_\infty}{\eta_N} &\approx (1 + 10^{19}\eta_N)^{-1} \approx 0.05 . \end{aligned} \tag{7.3.16}$$

Hence, in the limit $t \to \infty$, annihilation further decreases the nucleon-to photon ratio to the value $\eta_\infty \approx 10^{-19}$. It is easily seen from (7.3.16) that this value is obtained early on (say, for $\tau \approx 10$ or $t \approx 0.02$ s).

At this time, photons and $e^\pm$ pairs constitute $\sim$60% of all of the light particles. The $e^\pm$ pairs eventually annihilate into photons. Based upon this, today's ratio of the number of nucleons or antinucleons to that of photons in the charge symmetric Universe would be less than 10^{-19}. A 2.7 K RR temperature corresponds to a photon number density of about 400 cm^{-3}, resulting in $n_N = n_{\bar{N}} \lesssim 10^{-17}$ cm^{-3}, or $\rho_N = \rho_{\bar{N}} \lesssim 10^{-41}$ g cm^{-3}. These small values are arguments against the hypothesis of a charge-symmetric Universe. However, before discussing these conclusions, let us return to the calculation and pose a question. By what physical means is such a small dimensionless number, $n_N/n_\gamma \lesssim 10^{-19}$, obtained? Or, from what dimensionless parameters of the problem can one straightforwardly, even if only in order of magnitude, obtain this number? Even after obtaining a numerical result by a more precise method, an answer to this question is of inestimable value. For only in this way can one clearly understand the meaning of the result and the regime of its applicability.

Thus, we must determine the density of antinucleons after the "freeze-out." In the course of the expansion, as the temperature falls, the equilibrium density of antinucleons falls exponentially once $kT \lesssim Mc^2$. Clearly, by virtue of this exponential dependence on the temperature for $kT \leq Mc^2$, the time t_N of the "freeze-out" is roughly determined by the condition $kT_N = Mc^2/\alpha$, where α is a number of the order of, though greater than,

and $\alpha \gg 1$, as was shown above, $\alpha \sim 44$ in reality. In accordance with the behavior of the solution of a transcendental equation, α is the logarithm of large quantity.

It is clear that by this method one can only make a rough determination of T_N and consequently of t_N also. The idea is that, in a rough calculation, one can replace α by unity only in those places where α occurs as a factor. In the expression $n_{eq} \sim \exp(-\alpha)$, however, one can by no means consider $\alpha \sim 1$, since the exponential is very sensitive to α.

After the time of the "freeze-out" is found roughly, one can make use of the condition of equality of the time for the establishment of equilibrium and the dynamical time ($\sigma_0 c n_N t_N \approx 1$) to determine $n_N \approx n_{eq,N}$ at this instant. Let us make an order-of-magnitude estimate. The cross section is given by the expression $\sigma_0 \sim (\hbar/Mc)^2$. Moreover, inasmuch as the mass of nucleons and the mass of mesons are of the same order, the quantity $\hbar/Mc$—the Compton wavelength of the nucleon—also characterizes, in order of magnitude, the range of nuclear forces. We finally express the dynamical or expansion time t_N through α by making use of the relations

$$\rho_N \approx \frac{1}{G t_N^2} \approx \frac{kT_N}{c^2}\left(\frac{kT_N}{c\hbar}\right)^3 \approx M\left(\frac{Mc}{\hbar}\right)^3 \frac{1}{\alpha^4}.$$

The result is

$$t_N \approx \alpha^2 G^{-1/2} M^{-2} \hbar^{3/2} c^{-3/2}. \tag{7.3.17}$$

Now we substitute these expressions (with $\alpha = 1$) into the criterion characterizing the freeze-out: $\sigma v n_N t_N = 1$, to obtain

$$\left(\frac{\hbar}{Mc}\right)^2 c n_N G^{-1/2} M^{-2} \hbar^{3/2} c^{-3/2} \approx 1;$$
$$n_N \approx G^{1/2} M^4 c^{5/2} \hbar^{-7/2} \approx \left(\frac{GM^2}{\hbar c}\right)^{1/2}\left(\frac{Mc}{\hbar}\right)^3. \tag{7.3.18}$$

This expression may be compared with the expression for the number of photons,

$$n_\gamma \approx \left(\frac{kT_N}{c\hbar}\right)^3 \approx \left(\frac{Mc}{\hbar}\right)^3, \tag{7.3.19}$$

to find

$$\eta_N = \frac{n_N}{n_\gamma} \approx \left(\frac{GM^2}{\hbar c}\right)^{1/2} \approx \frac{M}{M_{Pl}}. \tag{7.3.20}$$

Hence, the "freeze-out" ratio η_N is nothing more than the ratio of the nucleon mass ($\sim 10^{-24}$ g) to the "Planck" mass [$(\hbar c/G)^{1/2} \approx 10^{-5}$ g]. Alternatively, the ratio $GM^2/\hbar c$ may be thought of as characterizing the strength of the gravitational interaction between nucleons (the analog of $e^2/\hbar c$ for the electromagnetic interaction). For nucleons, $GM^2/\hbar c \approx (M/$

$M_{\mathrm{Pl}})^2 \approx 10^{-38}$, so that $\eta_N \approx 10^{-19}$ is in remarkable agreement with the result of a more detailed calculation.

Hence, to the question of why so few nucleons and antinucleons remain in the hot, charge-symmetric model, one can give a succinct answer: this is due to the small size of their gravitational interaction.

In more detail, the number of remaining nucleon-antinucleon pairs depends on the competition between the rate of their annihilation and the expansion rate. The annihilation depends on the nuclear interaction, whose strength is of order unity, so that $\sigma_0 \approx (\hbar/Mc)^2$. The expansion is specified in such a form that the kinetic energy of the expansion equals the potential energy of the gravitational interaction (at the early stages this is always so; see § 1.8). As a consequence of this condition the expansion rate varies as $G^{-1/2}$. But just this quantity $G^{-1/2}$ enters into the expression for n_N/n_γ.

The calculation of the N-$\bar{N}$ annihilation that occurs during the expansion, a calculation set forth above, was apparently first performed by one of the authors (Zel'dovich 1965*d*). However, a numerical error, corrected above, crept into that 1965 result. The calculation without errors was presented in connection with the problem of the annihilation of quarks (see the following section) in the study of Zel'dovich, Okun', and Pikel'ner (1965). Finally, analogous results for antinucleons were obtained by Chiu (1966).

Let us return to the cosmological considerations. The question of charge-symmetric models has been discussed repeatedly in the literature. An evident difficulty, which we completely ignored above, is that the annihilation of nucleons and antinucleons occurs while gas is condensing to form stars and other celestial bodies. In connection with this, the efforts of a group of authors (Klein 1961; Alfvén 1965) have been directed to the invention of quite clever mechanisms for separating the protons and antiprotons in a gas cloud under the joint action of gravitational and magnetic fields.

From the calculations performed above, it is apparent that in a homogeneous model, annihilation occurs at an extraordinarily early stage, long before the formation of separate gas clouds. At this stage, the density of nucleons is already too small to be compatible with the density of ordinary matter observed today, so no subsequent separation can reconcile the charge-symmetric model with observations. The possibility that particles and antiparticles separate at the very early stage when $T \approx 300$ MeV is examined by Omnès; the principles were examined by us in the preceding chapter.

Ideally, one can conceive of a Universe that is, on account of its inhomogeneity, charge-symmetric only on the average. For the initial singular state, one assumes that surpluses of nucleons occur in some regions and surpluses of antinucleons in others. In essence, the situation for a particular

...tion does not differ from the asymmetric model, the principal model ...mined in this book. If a "region" is larger than what is contained within ...the horizon, then differences are presently not even observable in principle. If the regions are smaller, then one can observe γ-rays which are created ...at the boundaries of the regions where annihilation occurs even today.

But a calculation of the annihilation for a critical-density intergalactic ...gives a quantity exceeding the observed background of γ-rays. This difficulty goes away if one believes that there is no intergalactic gas or that the magnetic fields at the boundary between the regions of matter and antimatter prevent the displacement of ionized gases. This side of the problem is discussed in the following chapter. If the annihilation occurs long before the formation of galaxies, the γ-rays lose their energy by Compton scattering, but the energy thereby injected into the plasma distorts the Planck spectrum.

It is hardly necessary to discuss here, more concretely, all of the possible variants of this symmetric model (see, for example, Alfvén 1965; for an interesting new review, see Alfvén 1971). Nevertheless, it is perhaps worth remarking merely that the model that is inhomogeneous, but charge-symmetric in the mean, is not esthetically pleasing in comparison with the model that is homogeneous and charge-asymmetric. We must emphasize that the charge-symmetry properties of particles do not require that charge symmetry hold for the number or density of particles in the cosmological expansion. The last chapters of this book are dedicated to a study of additional problems of charge symmetry.

For the charge-asymmetric model one can also pose a question regarding the residual density of antinucleons. By virtue of the expansion one also obtains here a finite value for the asymptotic result: a finite, nonzero ratio of antinucleons to photons for $t = \infty$. Thanks to the surplus of baryons the equation for $\bar{\eta} \equiv n_{\bar{N}}/n_\gamma$, is linear as we shall show. The ratio of antinucleons to photons is $\bar{\eta} = n_{\bar{N}}/n_\gamma$ and the ratio of nucleons to photons is $\eta = \bar{\eta} + \eta_0$, where η_0 is the ratio of "excess" nucleons to photons:

$$\eta_0 = \left(\frac{n_N}{n_\gamma}\right)_0 \approx 3 \times 10^{-8}\Omega_0 h_0^2 \,. \tag{7.3.21}$$

The equation for the time variation of $\bar{\eta}$ is

$$\frac{d\bar{\eta}}{dt} = -\sigma_0 c n_\gamma(\eta_0 + \bar{\eta})\bar{\eta} + \frac{\psi}{n_\gamma} \tag{7.3.22}$$

Now, in equilibrium (with the temperature constant) we have

$$\frac{d\bar{\eta}_{eq}}{dt} = 0 \,; \quad \frac{\psi}{n_\gamma} = \sigma_0 c n_\gamma(\eta_0 + \bar{\eta}_{eq})\bar{\eta}_{eq} \,. \tag{7.3.23}$$

In the above, $\bar{\eta}_{eq}$ is the equilibrium ratio of antinucleons to photons *in the presence of excess nucleons*. In terms of the previously introduced η_{eq}, which

is the equilibrium ratio in the charge-symmetric case, $\bar{\eta}_{eq} \approx \eta_{eq}^2/\eta_0$. In the limit which concerns us ($\eta_0 \gg \eta_{eq}$, $\eta_0 \gg \bar{\eta}$), (7.3.22) becomes

$$\frac{d\bar{\eta}}{dt} \approx -\sigma_0 c n_\gamma \eta_0 \left(\bar{\eta} - \frac{\eta_{eq}^2}{\eta_0}\right). \tag{7.3.24}$$

As before, we estimate the critical time t_N, past which equilibrium will no longer be maintained:

$$\left[-\frac{1}{\bar{\eta}_{eq}}\left(\frac{d\bar{\eta}_{eq}}{dt}\right)\right]_N^{-1} = \theta_N t_N \approx (\sigma_0 c n_{\gamma,N}\eta_0)^{-1}. \tag{7.3.25}$$

Expressing $n_{\gamma,N}$ and t_N in terms of θ_N, we obtain

$$\begin{aligned} \theta_N &= \frac{kT_N}{Mc^2} \approx 10^{-6}(\Omega_0 h_0^2)^{-1/2}, \\ t_N &\approx 10^{-6}\theta_N^{-2} \approx 10^6(\Omega_0 h_0^2)\ \text{s}, \\ \bar{\eta}_{eq,N} &\approx e^{-2/\theta_N} \approx \exp\left[-10^{6\sqrt{(\Omega_0 h_0^2)}}\right]. \end{aligned} \tag{7.3.26}$$

In a Universe with a homogeneously distributed surplus of baryons, the equilibrium is maintained to very low temperatures. The corresponding density of antinucleons is improbably small—less than one within the horizon. Mathematically speaking, this occurs because the surplus does not depend on temperature, while the temperature is found from an algebraic equation not containing $\exp(-1/\theta)$. Only afterwards is the temperature substituted into the exponential.

The physical result is that the number of primordial antibaryons in the charge-asymmetric model is negligibly small: after 0.01 s, $T \lesssim 10$ MeV, $\theta \lesssim 0.01$, and $\bar{\eta}_{eq} \lesssim 10^{-73}$. In our epoch, antibaryons exist that arise from the interaction of cosmic rays with ordinary matter. Though their density is small, about 10^{-22} cm^{-3}, this density is significantly higher than that of any primordial antibaryons.

7.4 Relic Quarks in the Hot Model

As in the discussion of § 6.3, let us assume that there exist particles (quarks) with baryonic charge $\frac{1}{3}$ and fractional electric charge. It follows that the transformation of some kinds of quarks into others by means of β-decay is possible. Thus, assume that the quark denoted by q_1 is stable, i.e., that its mass is less than the masses of the quarks q_2 and q_3. Correspondingly stable in such a case is the antiquark $\bar{q}_1$.

We emphasize that up to now quarks have been observed neither as constituents of primary cosmic rays, nor among the products of the interaction of cosmic rays with nuclei, nor as free quarks at accelerators. However, the hot model of the Universe leads to a conclusion about the possible occurrence of a noticeable density of relic quarks—free or attached to nuclei. In several countries, therefore, programs have been initiated to

...ept an observation of such quarks in ordinary matter by means of ... ments of the same type as the classic experiment of Millikan for ...ining the charge on the electron. In connection with this, an analysis ...f the equilibrium and freezing-out of quarks in the hot model is of astro... ...al interest. This, of course, assumes that the hypothesis of their ...ence is true. If quarks were observed, then their density and distribu... ...n the Universe would give totally new, quite valuable information ...out the evolution of the Universe. In sum, though, their existence is ...ps a highly attractive, but not proved. Indeed, the existence of ... as real particles is more and more subject to doubt. And cosmology ...ed no small role in the development of these doubts.

...Thus, calling the mass of a quark m_q, we must expect that for $kT/m_qc^2 > 1$ ...the equilibrium density of quarks and antiquarks does not differ from the ...density of all other particles of similar or lower mass. As the temperature ...falls during the expansion, the density decreases as exp $(-m_qc^2/kT)$. The ...full theory of the freezing-out of quarks is built on the totally analogous ...theory of the freezing-out of nucleons and antinucleons in the symmetric ...model (see the previous section).

We assume that the mass of a quark is several times larger than the mass ...of a nucleon. Therefore, the equilibrium density soon becomes insignifi... ...antly small—less than one quark in the whole observable Universe—at a ...temperature of the order of several kilovolts. Long before this, however, ...the rate of the reactions leading to the establishment of equilibrium be... ...comes too small, and the "freezing out" of the remaining q and $\bar{q}$ occurs. ...Thereafter their density continues to decrease simply because of the general ...expansion. At the same time, the ratios of their densities to that of nucleons, ...i.e., n_q/n and $n_{\bar{q}}/n$, tend to a constant.

Quarks interact strongly with one another. The low reaction rate is ...connected with the necessity of always having at least two quarks or ...of having a quark-antiquark pair as reactants. Since a baryon B consists of ...three quarks, it might at first glance seem that a three-body collision is ...necessary. In fact, though, the process

$$q + q \rightarrow B + \bar{q} \tag{7.4.1}$$

...is energetically quite possible since $m_q > m_B$. And two-body collisions ...clearly occur more frequently than do three-body ones. For example, a q-$\bar{q}$ ...pair annihilates and transforms into mesons. The inverse of the characteristic ...time for this reaction that "does away" with quarks is given by the ex... ...pression

$$\tau^{-1} = \sigma v n_q . \tag{7.4.2}$$

...Here σ, the cross section for the reaction, is of the order of $(\hbar/m_qc)^2c/v$; ...v is the velocity of the quarks; and n_q or $n_{\bar{q}}$ is the density of the corresponding ...particles.

The freezing-out of the quarks occurs at a temperature much higher than

that characterizing the antinucleons. It occurs during the period of approximate charge symmetry with respect to N and $\bar{N}$ even in the charge asymmetric model. With great accuracy, therefore, $n_q \approx n_{\bar{q}}$ or $n_q - n_{\bar{q}} \ll n_q$. Reaction (7.4.1) promotes the further decrease of the difference $n_q - n_{\bar{q}}$. Keeping this inference (which is interesting by itself and will, we hope, eventually be confronted with observations) in mind, we will in the subsequent development simply write n_q. Hence, the reaction rate tends to zero as n_q decreases.

Recall the results of the preceding section. To locate the time at which the freezing-out occurs, one must compare the reaction rate with either the expansion rate or the rate of change of the equilibrium density (the difference between the corresponding characteristic times is not very large). Thus, there exists a certain $n_{q,f}$ and a certain t_f, for which

$$\tau_f = t_f ,$$

where t_f is the dynamical time. Simply speaking, t_f is the time elapsed from the time when $\rho = \infty$. After this, the corresponding ratio n_q/n is approximately (up to a numerical coefficient—accurately) conserved in the course of the further expansion.

The freezing-out phenomenon selects a definite value for $n_{q,f}$. The equilibrium density $n_{q,\mathrm{eq}}$ depends exponentially on the mass of the quarks. But, if $n_{q,f} = n_{q,\mathrm{eq}}$ is specified, then the freezing-out time t_f and the temperature T_f are themselves adjusted to the value of the mass of the quark m_q via $kT_f = m_q c^2/\ln \alpha$. Here α depends on the cross section σ, on the expansion rate, and on other factors, but the dependence on these factors is only logarithmic. As a result, the value of $n_{q,f}$ obtained does not depend exponentially on m_q—a most important result. It is for this reason that one can hope to observe relic quarks—even if their mass is large and observation of them by means of the methods of particle physics is not feasible.

Repeating now the reasoning of the preceding section, we are led to the formula

$$\frac{n_{q,f}}{n_{\mathrm{tot}\,f}} \approx \left(\frac{G m_q^2}{\hbar c}\right)^{1/2} \approx 10^{-19}\left(\frac{m_q}{M}\right). \tag{7.4.3}$$

Here n_{tot} is the density of all species of particles at the freeze-out time; the numerical value used for m_q is of the order of a nucleon mass. When discussing the entropy of the hot model, we found

$$\frac{n}{n_{\mathrm{tot}}} \approx S^{-1} = \text{const.} \approx 10^{-9}, \tag{7.4.4}$$

where n is the density of baryons. From this we find

$$\frac{n_{q,f}}{n_f} \approx \frac{10^{-19}}{10^{-9}} = 10^{-10}. \tag{7.4.5}$$

Hence the ratio of relic quarks to baryons is proportional to the specific entropy. The hot model predicts that there are $\sim 10^{-10}$ quarks and as many

...quark for each baryon. The calculation lays no claim to accuracy, but ... result is impressive. Indeed, the abundance of gold is (Au/H) $\approx 10^{-12}$, ... that of radium (Ra/H) $\approx 10^{-18}$. Hence, if quarks exist, then in ... and certainly in the Earth's crust—they must be encountered ... often than is gold!

These results relating to the freeze out of quarks were obtained by ... Okun', and Pikel'ner in 1965. When this work was carried out, ... validity of the hot model was still not considered proved. The results ... in that work therefore serve as upper limits when applied to the hot ... and lower limits when applied to the cold model. Also examined in ... study is the question of the subsequent fate of quarks—in particular, ... the possibilities of their sticking to nuclei and of their combustion in ...

... notice that, by virtue of the charge symmetry, there always remain q ... $\bar{q}$ having both signs of charge. If q is stable and has a charge $Z = +2e/3$, ... then relic $\bar{q}$ with charge $Z = -2e/3$ also remains. In the alternate case, ... there remains q with charge $Z = -e/3$ and $\bar{q}$ with charge $Z = e/3$.

At first glance $n_q = n_{\bar{q}}$ facilitates the further combustion of q and $\bar{q}$ in ..., for there is no Coulomb barrier to hinder a collision. In fact, however, ... was noticed in the work of Zel'dovich, Okun', and Pikel'ner (1965), ... negative q or $\bar{q}$ are always stabilized by adhering to nuclei. Therefore, the ... presence of particles with charges of both signs promotes the survival of ... negative q (or $\bar{q}$) in stars. Apparently, the combustion is not very significant.

The question of the distribution of quarks during the periods of stellar ... evolution and of planetary formation is very difficult. We shall not pursue it ... here.

We once again emphasize that quarks have never been found, and that their existence as separate particles—not, for example, as excitations, analogous to phonons in a solid body—is subject to strong doubt. Recall, however, the remark in § 5.3 concerning the partons or quarks with integral charge. It would be more difficult to observe them than to observe fractionally charged quarks. At the present time it is not clear whether experiment will demonstrate the absence in nature of neutral or integrally charged quarks in the quantity predicted by theory. Concerning the question of experimental methods relevant to the searches for relic quarks, we merely indicate some of the appropriate articles: Bechi, Gallinaro, and Morpurgo (1965); Braginsky (1966); Chupka, Schiffer, and Stevens (1966); Braginsky, Zel'dovich, Martinov, and Migulin (1967, 1968); and Pikel'ner and Vainshteyn (1966).

7.5 Nucleosynthesis in the Hot Model

In this section we consider the nuclear reactions which determine the chemical composition of the prestellar matter. We examine these processes

in the charge-asymmetric model, which in our opinion correctly describes the observed Universe. In the course of the cosmological expansion, after the temperature falls to a value $T < Mc^2/k$, only nucleons remain (and not antinucleons). The ratio of nucleons to light particles is $\sim 10^{-8}$. The nucleons subsequently participate in nuclear reactions. As we shall see, these reactions terminate roughly 100 s after the beginning of the expansion, at which time the temperature is several tens of kilovolts or $\sim 3 \times 10^8$ K.

Afterwards, the chemical composition remains unchanged. From this matter, stars of the first generation form—presumably at a time corresponding to $z \approx 10$ or $z \approx 5$, i.e., in a rather recent era. Only after the subsequent compression of the matter in stars to high density and temperature does a new stage of nuclear reactions begin.

The first attempts to compute the primordial chemical composition of the Universe were based on the idea of thermodynamic equilibrium. We note here the pioneering role of Pokrovsky (1928). The mere formulation of the question about the conditions required for the formation of the elements deserves great praise. This idea represented a bold application to cosmology of achievements of nuclear physics.

However, it was soon explained that one need not choose a density and temperature for the matter in such a way that the observed composition of the Universe would be an equilibrium composition. To Gamow (1946) belongs the credit for an evolutionary formulation of the problem. In the expanding Universe, the kinetics of the various processes is essential.

We saw above (see chap. 6) that equilibrium is maintained for fast processes. So why do the nuclear reactions turn out to be slow? Why is kinetics essential for a description of them? The point is that nuclear processes consist of (*a*) transformations of neutrons into protons and the reverse, processes that are connected with the emission and capture of leptons (e^+, e^-, ν_e, $\bar{\nu}_e$) and therefore with the weak interaction, and (*b*) binary nuclear reactions of the type $n + p \rightarrow \mathrm{D} + \gamma$, $\mathrm{D} + \mathrm{D} \rightarrow {}^3\mathrm{He} + n$, and so forth, which proceed relatively slowly because the ratio of the baryon density to the photon density is small.

As was already noted in chapter 5, the early studies of the problem of nucleosynthesis abound with errors. Originally, Gamow (1946) assumed that at a very high density, nucleons are all neutrons. Subsequently the neutrons decay. In unpublished work, Fermi and Turkevich examined the nuclear reactions whereby neutrons combine with protons and also subsequent reactions. Later, Hayashi (1950) noticed that at a high temperature, one higher than 2 MeV, or 2×10^{10} K, electrons, positrons, neutrinos, and antineutrinos cause the rapid transformation of neutrons into protons, and vice versa, according to the reactions:

$$\begin{aligned} e^+ + n &\rightleftarrows p + \bar{\nu}\,, \\ \nu + n &\rightleftarrows p + e^-\,. \end{aligned} \tag{7.5.1}$$

and electron neutrinos are considered subsequently, so we omit the subscript e.) Independent of the choice of the initial state, therefore, thermodynamic equilibrium is established at a high temperature so that

$$\frac{n_n}{n_p} = \exp\left(\frac{-\Delta mc^2}{kT}\right), \tag{7.5.2}$$

here Δm is the neutron-proton mass difference:

$$\Delta mc^2 = 1.29 \text{ MeV}.$$

There are, of course, no complex nuclei at all initially. So, for example, at the time $t = 0.01$ s, when the temperature equals 9.5 MeV, the time required for the establishment of equilibrium is of the order of 2×10^{-3} s, and the equilibrium abundance relation is $n_n/n_p = 0.47/0.53$.

As the temperature drops, the equilibrium ratio falls, but the rate of processes (7.5.1) also decreases. As with other processes that slow down at some stage of the expansion (see § 7.1–4), these break up into two stages. During the first stage, $n_n = n_{n,\text{eq}}$, with $n_{n,\text{eq}}$ denoting the equilibrium density. At the end of the first stage, a "freeze-out" occurs; the reactions are then too slow to maintain equilibrium. During the second stage, the remaining neutrons slowly decay with a half-life of 11 minutes known from laboratory experiments. (The modern value for the average lifetime of the neutron is 935 ± 14 s. The half-life is obtained from this upon multiplication by ln 2.)

Let us now present a quantitative calculation. Hayashi's (1950) calculations were extraordinarily complicated, and the premises upon which they were based do not completely correspond to present knowledge of the weak processes. These calculations were repeated by V. M. Yakubov and set forth in the work of Zel'dovich (1965*d*). Still later appeared the work of Peebles (1966*a*, *b*); of Wagoner, Fowler, and Hoyle (1967); of Wagoner (1973); and of Reeves, Audouze, Fowler and Schramm (1973).

Thus suppose that the densities n_n and n_p satisfy the equations

$$\begin{aligned} \frac{dn_n}{dt} &= -a(t)n_n + b(t)n_p, \\ \frac{dn_p}{dt} &= a(t)n_n - b(t)n_p. \end{aligned} \tag{7.5.3}$$

The coefficients $a(t)$ and $b(t)$ depend on the temperature and hence on the time through the dependence of the time on the temperature. Since the number of nucleons is insignificant in comparison with the number of light particles, the reactions involving neutrons and protons have virtually no influence on the densities of e^-, e^+, ν, and $\bar{\nu}$; hence these reactions do not affect the functions $a(t)$ and $b(t)$. For this reason the equations are linear.

For a calculation of $a(t)$ and $b(t)$, one must take account of the Pauli principle, even though one can neglect the motion of protons and neutrons.

The distributions of e^-, e^+, ν, and $\bar{\nu}$ in momentum space are given by the well-known expression of Fermi statistics:

$$g_e = \frac{1}{\exp(-E_e/kT) + 1}, \quad g_\nu = \frac{1}{\exp(-E_\nu/kT) + 1}; \qquad E_e = (p_e^2c^2 + m_e^2c^4)^{1/2}, \quad E_\nu = cp_\nu. \tag{7.5.4}$$

The rate at which $p \rightarrow n$ is

$$b = A[\int\int g_{\bar{\nu}}[1 - g_{e^+}]\delta(E_{\bar{\nu}} - \Delta mc^2 - E_{e^+})d^3p_{\bar{\nu}}d^3p_{e^+} + \int\int g_{e^-}[1 - g_\nu]\delta(E_{e^-} - \Delta mc^2 - E_\nu)d^3p_{e^-}d^3p_\nu]. \tag{7.5.5}$$

The first term describes the reaction $\bar{\nu} + p \rightarrow n + e^+$; the second, $e^- + p \rightarrow n + \nu$. The factors in square brackets, $[1 - g]$, take account of the vacancies in momentum space: the reaction is forbidden if the final state of e^+ or of $\bar{\nu}$ is already occupied. The delta function takes account of energy conservation. Only electrons and antineutrinos with sufficiently large energy ($E_{e^-} > \Delta mc^2$; $E_{\bar{\nu}} > \Delta mc^2 + m_ec^2$) can initiate the reactions. Thanks to the delta function, the sixfold integral contracts to a fivefold integral. In the limit when the energies considered are far greater than m_ec^2 and Δmc^2, the integral is proportional to E^5. The constant A is proportional to the square of the weak-interaction constant g_{wk}. One can, by the way, express A through the decay probability of a free neutron.

The thermodynamic identity,

$$b(t) = a(t)\exp\left(-\frac{\Delta mc^2}{kT}\right) \approx a(t)\exp(-1.4t^{1/2}), \tag{7.5.6}$$

holds if all processes are accounted for. For high temperatures, or for small t,

$$a \approx b \sim T^5 \sim t^{-5/2}, \quad a \approx b \approx 4 \times 10^{-6}T_9^5\ \mathrm{s}^{-1}; \tag{7.5.7}$$

here $T_9 = T/10^9\ \mathrm{K}$. For low temperatures, or for $t \gg 1$ s,

$$a(t) \rightarrow W \approx 10^{-3}\ \mathrm{s}^{-1}, \quad b(t) \rightarrow 0, \tag{7.5.8}$$

where W is the decay probability for a free neutron. For the neutron abundance by mass: $X_n \equiv n_n/(n_n + n_p)$ we have

$$\frac{dX_n}{dt} = b - (a + b)X_n. \tag{7.5.9}$$

The solution to (7.5.9) is

$$X_n(t) = \int_0^t b(\tau)\exp\left[-\int_\tau^t (a + b)d\tau'\right]d\tau. \tag{7.5.10}$$

We now introduce the following notation:

$$a + b = f + W.$$

[illegible] large t the function $f \to 0$ since $a \to W$ and $b \to 0$. The asymptotic [illegible]ion for $X_n(t)$ when $t \gg 1$ s thus has the form

$$X_n(t) = X_{n*} e^{-Wt}\,, \qquad (7.5.11)$$

$$X_{n*} = \int_0^t b(\tau) \exp\left[W\tau - \int_\tau^t f d\tau'\right] d\tau\,,$$

[illegible]re the characteristic quantity X_{n*} is given by the integral in the second [illegible]quality. The integrand falls sharply for $t > 1$ s and grows again only for [illegible]10^4 s. Therefore, the definition does not depend on the specific choice of [illegible]d $100 < t < 10^4$ s. For $t \to 0$, $b(t) \to \infty$ and $f(t) \to \infty$ though the exponential goes to zero faster than does $b(t)$. The integrand as a whole tends to zero as t tends to zero.

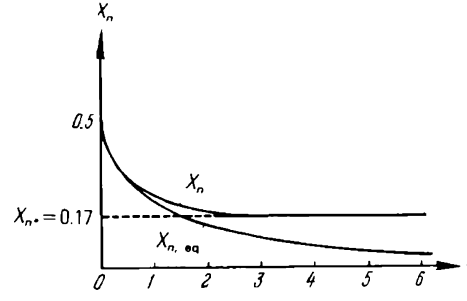


FIG. 31.—Evolution of the neutron abundance in the hot model. The curve labeled X_n gives the actual abundance, while that labeled $X_{n,eq}$ gives the equilibrium abundance.

The curves of $X_n(t)$ and $X_{n,eq}(t)$ are shown in Figure 31. The quantity X_{n*} represents the neutron mass fraction after freeze-out occurs. The value of X_{n*} calculated in this way turns out to be 0.165; i.e., it differs little from the result of Hayashi (1950).

One can also make a rough estimate of X_{n*} without such detailed calculations by using the same method to which we repeatedly resorted in the preceding sections. Indeed, reactions halt and there occurs a freeze-out whenever the condition (see eq. 6.2.8)

$$t = \tau \qquad (7.5.12)$$

is satisfied, where t is the time from the beginning of the expansion (6.2.3) and τ^{-1} the reaction rate. In our case, this time is

$$\tau \approx 10^5 T_9^{-5}\ \text{s}\,, \qquad (7.5.13)$$

in accordance with equation (7.5.7). Substituting expressions (7.5.13) and (6.2.3) into equation (7.5.12) enables us to find the temperature at which the freeze-out occurs. We take the coefficient κ to be of order ~ 4. This value is obtained by taking account of only the known kinds of elementary

particles and by assuming that hypothesized, unknown particles do not exist (see more about this at the end of the section). We then obtain

$$T_* \approx 10^{10}\,\mathrm{K}\,, \qquad kT_* \approx 0.9\,\mathrm{MeV}\,, \tag{7.5.14}$$

for the temperature when the freeze-out occurs. Now, substituting this temperature T_* into equation (7.5.2), we find the ratio n_n/n_p at freeze out

$$\left(\frac{n_n}{n_p}\right)_* = e^{-\Delta mc^2/kT_*} \approx e^{-1.5} \approx 0.2\,. \tag{7.5.15}$$

This yields

$$X_{n*} = \left(\frac{n_n}{n_n + n_p}\right)_* \approx 0.17\,, \tag{7.5.16}$$

in agreement with the result of the exact calculation.

Subsequently, when $t \approx 100$ s, neutrons and protons join to form deuterons,

$$n + p \rightarrow \mathrm{D} + \gamma\,, \tag{7.5.17}$$

after which a chain of reactions follows:

$$\mathrm{D} + \mathrm{D} \begin{cases} \rightarrow {}^3\mathrm{H} + p\,; & {}^3\mathrm{H} + \mathrm{D} \rightarrow {}^4\mathrm{He} + n\,. \\ \rightarrow {}^3\mathrm{He} + n\,; & {}^3\mathrm{He} + n \rightarrow {}^3\mathrm{H} + p\,. \end{cases} \tag{7.5.18}$$

As a result of this, ^{4}He is formed.

Deuterium represents an "entropy barrier" on the path to the formation of ^{4}He. At a temperature higher than 100 keV, the equilibrium density of deuterium is negligible. This is because photodisintegration $\gamma + \mathrm{D} \rightarrow p + n$ successfully competes with deuterium formation. But ^{4}He formation requires a sequence of two-body collisions. (It is curious that the direct process $\mathrm{D} + \mathrm{D} \rightarrow {}^4\mathrm{He} + \gamma$ has a totally insignificant cross section because the process is forbidden in several ways.) The abundance of ^{4}He therefore is greater the higher the density of baryons at a given temperature, i.e., the lower the specific entropy. For a low entropy, all of the neutrons will be transformed into ^{4}He; consequently by mass, $\sim$33% ^{4}He and $\sim$67% hydrogen result. Note that the ratio T^3/ρ, where ρ is the density of ordinary matter, is a measure of the entropy.

In Figure 32 are presented the results of the calculations of Wagoner, Fowler, and Hoyle (1967; see also Wagoner 1973) for an RR temperature $T = 2.7$ K and for different values of the present-day density ρ_0, to account for the unreliability with which the true value of ρ_0 is known. For the calculation of the abundance of ^{3}He and ^{7}Li, they allowed for the conversion of ^{3}H and ^{7}Be into ^{3}He and ^{7}Li, respectively, via β-decay.

Figure 33 gives a general picture of the time dependence of the abundances of p, n, D, and ^{4}He in the hot model for $T^3/\rho \approx 3 \times 10^{30}$, i.e., $T_0 = 2.7$ K and $\rho_0 \approx 7 \times 10^{-30}\,\mathrm{g\,cm^{-3}}$); the calculations are those of Doroshkevich

ed survey (1969). The question of the actual abundance of helium and other light elements in the prestellar matter is addressed in the next section.

Let us here recapitulate. What are the premises of and possible weak points in the calculation? First, the matter density ρ_0, or the entropy

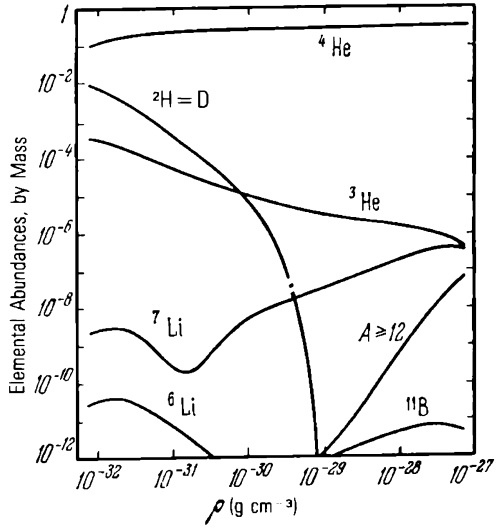


FIG. 32.—The relative abundance of the chemical elements synthesized at the beginning of the cosmological expansion. Plotted horizontally is the mean matter density of the Universe today.

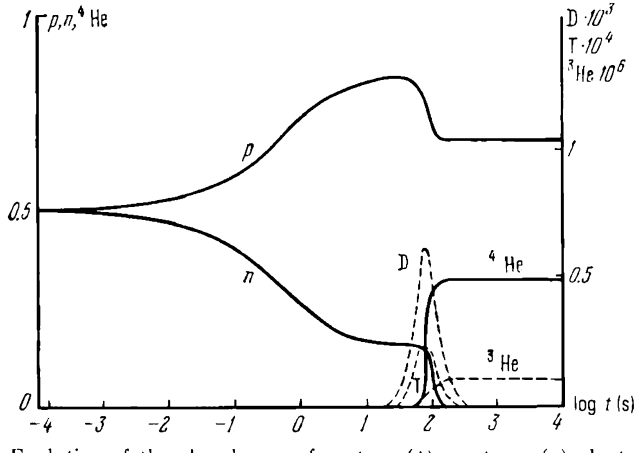


FIG. 33.—Evolution of the abundances of protons (p), neutrons (n), deuterium (D), helium-4 (^{4}He), helium-3 (^{3}He), and tritium (T) in the hot model as a consequence of nuclear reactions. The vertical scale on the left applies only to p, n, and ^{3}He; that on the right, only to D, T, and ^{3}He.

$\sim T_0^3/\rho_0$, are not reliably known. The quantity T^3/ρ hardly changes with time; however, it is possible for ρ_0 to be anything from 3×10^{-31} g cm^{-3} to 3×10^{-29} g cm^{-3}. Note that the rates for the weak interactions $n \rightleftarrows p$ and the freeze-out time do not depend on ρ_0/T_0^3. Subsequent nuclear reactions do depend on ρ_0/T_0^3. For the range of densities of interest, viz., 3×10^{-31} g cm$^{-3} < \rho_0 < 3 \times 10^{-29}$ g cm^{-3}, however, this dependence is not important in determining the ^{4}He abundance. Indeed, because of this the predicted values of the abundance by mass of ^{4}He only range between 25% and 29%. The amounts of D, ^{3}He, and ^{7}Li depend very strongly on the ratio ρ_0/T_0^3; the absolute amount of each is very small. An account of their abundance in celestial bodies and of their synthesis at present is set forth in the next section.

A summary of the results of the research on the cosmologically important nuclear reactions presented above is then as follows: Very roughly, one can say that all of the neutrons that survive till the freeze-out are converted into ^{4}He during the period of nucleosynthesis, $t \approx 10$–100 s. Hence, the abundance by mass, Y, of ^{4}He in the prestellar matter in a first approximation is twice the abundance by mass of the neutrons at freeze-out:

$$Y \approx 2X_{n*} . \tag{7.5.19}$$

Substituting here the value $X_n^* \approx 0.165$ obtained above, we find

$$Y \approx 0.33 . \tag{7.5.20}$$

This is in essence the maximum amount of helium that can be obtained. The actual value of Y is somewhat less, as is apparent from Figure 33, but formula (7.5.20) is sufficiently good for a reasonable estimate. Thus $X_n^* [n_n/(n_n + n_p)]_*$, evaluated at freeze-out, has a decisive influence on the outcome of nucleosynthesis.

To what degree is the value of X_n^* stable with respect to changes in the various parameters of the model? Above it was noted that X_n^* does not depend on ρ_0/T_0^3 since equations (7.5.3) are linear. Hence, one can neglect the interactions of neutrons and protons with one another for $T > 1$ MeV. But X_n^* does depend on the expansion rate or, more precisely, on the change of the temperature with time. This is immediately apparent from formula (7.5.12). On the left stands $t(T)$, the time from the beginning of the expansion, which coincides with the time for a change in the parameters of the model. Earlier, for a determination of the temperature at which the freeze-out occurs, we used formula (6.2.3) with the coefficient $\kappa \approx 4$. This coefficient depends on the number of kinds of elementary particles present at a given temperature. Of course, the coefficient κ only takes account of known particles. Shvartsman (1969) indicated that an assumption of a strong increase in κ changes the conclusions about the chemical composition of the primordial matter. Such an increase is in principle possible if there are many unknown, weakly (or superweakly) interacting, and therefore

undiscovered, particles. Such particles must have a small or zero rest mass. This is because as the Universe expands, the ratio of their density to the density of ordinary matter would not change so that the average density due to these particles would vastly exceed the density due to the known forms of matter. But this is impossible, as was seen in the discussion of the question of the mass of the muon neutrinos.

Massless particles such as neutrinos or gravitons cannot be excluded by such reasoning. We emphasize once again that the hypothesized particles do not participate directly in the reactions leading to the synthesis of the elements, but influence them indirectly by changing the dynamics of the expansion. For an estimate of the effect of these particles on the nuclear reactions, we make use of formulae (7.5.12) and (7.5.19). We shall denote by μ the ratio of the total density of matter, including the density of unknown particles, to the density of the known particles in thermodynamic equilibrium. In the difference $\rho_{\text{tot}} - \rho_{\text{known part}}$ we include not only the density of unknown, massless particles, but also the density of muon neutrinos and antineutrinos if there should exist a large excess of one or the other in comparison with their equilibrium density (i.e., neutrino degeneracy):

$$\mu \equiv \frac{\rho_{\text{tot}}}{\rho_{\text{known part}}} \approx \frac{\kappa}{\kappa_0}, \tag{7.5.21}$$

where $\kappa_0 \approx 4$ is the value of the known particles. With the help of equation (6.2.3), we find

$$t \approx \frac{10^2}{T_9^2 \mu^{1/2}} \text{ s}. \tag{7.5.22}$$

Then, using equations (7.5.12), (7.5.13), and (7.5.20), we obtain

$$T_* \approx 10^{10} \mu^{1/6}; \quad \left(\frac{n_n}{n_p}\right)_* \approx \exp\left(-1.5/\mu^{1/6}\right), \tag{7.5.23}$$

$$Y \approx 2X_{n*} = \frac{2(n_n/n_p)_*}{1 + (n_n/n_p)_*}. \tag{7.5.24}$$

The curve in Figure 34 shows Y as a function of μ.

It is possible to infer that $\mu > 4$ is impossible, since such a value gives $Y > 50\%$, which is clearly excluded by observations.

We emphasize that the reasoning presented gives much stricter limitations to μ than does a comparison of the present-day value of the Hubble parameter with the age of the Earth. The latter estimate only leads to the inequality $\mu < 10^5$ (Zel'dovich and Smorodinsky 1961).

Notice that a huge increase in the energy density of massless particles, $\mu \gg 1$, can lead to such an increase for the expansion rate that the reactions (7.5.17) and (7.5.18) responsible for the formation of ^{4}He during the second stage, may not be able to proceed. This would result in a low abundance of ^{4}He. But in this case, $\mu \gg 10^5$, the particles would also strongly influence

the present-day expansion rate. This, however, is incompatible with estimates of the age of the Earth and of the Hubble parameter.

A change in the parameter κ thus leads to a change in X_n^*. Now a possible surplus of electron neutrinos or antineutrinos can also affect the value of X_n^* significantly. Fowler (1970) turned his attention to this possibility. Up to now we have assumed that the total leptonic charge vanishes (i.e., we have supposed that the number of neutrinos equals the number of antineutrinos). If one refrains from such a requirement, then the excess of electron neutrinos leads, via reaction (7.5.1), to a change in the equilibrium ratio of neutrons to protons at the time of the freeze-out. Formula (7.5.2) will then be changed. A secondary effect is that the presence of the excess particles speeds the expansion and, consequently, changes the time of the freeze-out.

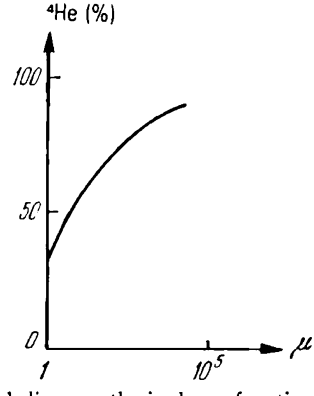


FIG. 34.—The amount of helium synthesized as a function of the number of unknown, weakly interacting particles.

In the course of the expansion, the chemical potential of the neutrinos varies linearly with temperature. We shall denote the ratio by $\phi = \mu_\nu/kT$. Then the following relations are approximately valid:

$$n_{\bar{\nu}} \sim T^3 e^{-\phi}\,; \qquad n_\nu \sim T^3(1+\phi)^3\,. \tag{7.5.25}$$

We thus have

$$\frac{n_{\bar{\nu}}}{n_\nu} = \frac{e^{-\phi}}{(1+\phi)^3}\,. \tag{7.5.26}$$

Instead of formula (7.5.2), we now have

$$\left(\frac{n_n}{n_p}\right)_{\text{eq}} = \exp\left(-\Delta mc^2/kT - \phi\right), \tag{7.5.27}$$

and instead of formula (7.5.15),

$$\left(\frac{n_n}{n_p}\right)_* = \exp\left(-\Delta mc^2/kT_* - \phi\right). \tag{7.5.28}$$

cally speaking, therefore, Y depends both on κ and on ϕ. If $\phi > 1.6$, Y is less than 30% for any κ!

In Figure 35 the dependence of Y on the specific leptonic charge $L_\nu = n_{\nu_e} - n_{\text{baryon}}$ is shown for $\kappa = 4$ and $\mu = 1$. A large excess of antineutrinos leads to $Y \approx 1$, contradicting observations; a large excess of neutrinos leads to $Y \approx 0$.

The assumption $\phi \neq 0$ also leads to the existence of an excess of neutrinos over antineutrinos today. However, although this excess is sufficient to cause changes in Y, it is of course still too small for observations of it to be made by any present day means.

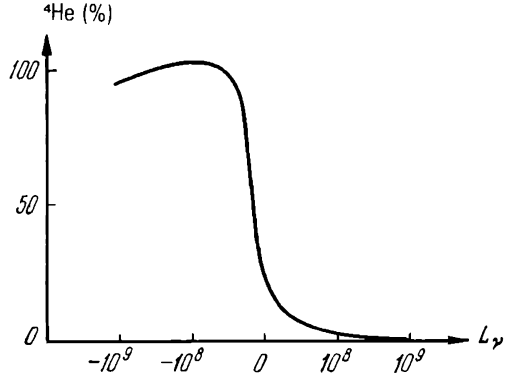


FIG. 35.—The amount of helium synthesized as a function of the specific lepton charge of the Universe.

7.6 A Comparison of the Observational Data on the Distribution of Light Elements in the Universe with the Predictions of Theory

What do the observations say about the distribution of chemical elements in the Universe? Do they confirm the predictions of the theory of the hot Universe?

First of all, it is clear that even knowing the chemical composition of the celestial bodies at the present time, one can still not compare these data directly with cosmological theory because chemical elements can be synthesized and also destroyed in the course of the evolution of celestial bodies. For example, this can occur either in stars or by the interaction of cosmic rays with the interstellar medium. It is therefore necessary to analyze the question of the time evolution of the distribution of the chemical elements; only after this can one attempt to compare the theory with the observations. We will present the data on the distribution of the chemical elements first, then analyze the possibility that elements are synthesized or destroyed in various processes, and, finally, draw conclusions about the

chemical composition of the prestellar matter out of which the first objects formed. A wonderful review of this question is given in the study by Reeves, Audouze, Fowler, and Schramm (1973), a study to which we will largely confine ourselves in the subsequent account (see also the detailed bibliography in this work).

We will be discussing only the light elements, for, as was repeatedly noted above, the synthesis of carbon and of heavier elements can be completely explained by processes occurring in the evolution of stars (see Truran and Cameron 1971).

Table 4 is taken from the study conducted by Reeves, Audouze, Fowler, and Schramm (1973). Appearing in it is the abundance by number of the light elements. In the preceding section, the abundance by mass was used.

We begin the discussion with ^{4}He, which is the second most prevalent element in the Universe. The observational aspects of the ^{4}He abundance are analyzed in the review by Searle and Sargent (1972). The amount of ^{4}He in young stars, in stellar atmospheres, and in the interstellar gas can be determined accurately. For example, the theory of stellar structure predicts definite relations among the mass, radius, and luminosity of young stars, and—what is important here—the helium content of the stars enters into these relations as a parameter. A comparison of theory with observations gives the ^{4}He abundance. The abundance of ^{4}He in stellar atmospheres and in the interstellar gas is determined using spectroscopic measurements and the theory of the excitation and ionization of ^{4}He.

With few exceptions (see below), it turns out that the abundance by mass of helium always lies in the interval

$$0.24 < Y < 0.32 . \qquad (7.6.1)$$

Recall the connection between the helium abundance by mass Y and the abundance by number ^{4}He/H. Here we use the fact that the remaining elements occur only in negligible amounts in the Universe. The above result also applies to the abundance of ^{4}He in the nearest galaxies, where helium is observed through its emission lines in ionized gas.

Now let us say something about the exceptions. The atmospheres of several stars exhibit an unusually small amount of helium. Such stars are extremely rare, and all show strong anomalies in their content not only of helium but also of other elements. Apparently, processes occurring on the surfaces of such stars sharply distort their surface chemical composition.

Could all of the observed ^{4}He be synthesized in the course of evolution of the first generation of stars, which would then return this helium-enriched gas to the interstellar medium so that from this helium "polluted" gas the present-day objects could develop at a much later stage? A careful analysis shows that this is hardly possible. A most convincing argument against such a possibility follows from an analysis of luminosity–spectral-class diagrams for old star clusters. These diagrams depend on the age of

TABLE 4

COSMIC ABUNDANCES

ELEMENT RATIOS	OBSERVED VALUE (by number)			
	Interstellar Gas	Stars	Sun	Planetary System
D/H	$(2\pm1)\times10^{-5}$	$<6\times10^{-4}$	$<5\times10^{-6}$	1.5×10^{-4}[a] $(2.9\text{–}7.5)\times10^{-5}$[b]
$^3He/^4He$	$\lesssim5\times10^{-4}$	...	$(4\pm2)\times10^{-4}$	$(1\pm0.5)\times10^{-4}$
$^4He/H$	0.11 ± 0.03	~0.10	$\sim0.08\text{–}0.10$	...
Li/H	$\sim10^{-9}$	$<10^{-12}\text{–}4\times10^{-9}$[c] 10^{-7}[d]	$\sim10^{-11}$	1.2×10^{-9}
$^7Li/^6Li$	...	>10	>20	12.5 ± 0.2
Be/H	$\sim10^{-9}$	$<10^{-12}\text{–}5\times10^{-11}$[c]	$\sim10^{-11}$	2×10^{-11}
B/H	...	...	$<3\times10^{-10}$	$3\times10^{-10}\text{–}3\times10^{-9}$
$^{11}B/^{10}B$	...	...	...	4 ± 0.5

[a] For the Earth and meteorites.

[b] For the atmosphere of Jupiter.

[c] This value varies sharply with the type and the age of stars.

[d] In several red giants.

the stars and on the *initial* helium content of the stars. It turns out that the initial ^{4}He content of the oldest (and therefore most metal poor) stars in our Galaxy is about $Y \approx 0.3$. This agrees with the results of Searle and Sargent (1972) and Reeves, Audouze, Fowler, and Schramm (1973), who come to the conclusion that the ^{4}He content of the prestellar matter is probably roughly the same as of the ^{4}He content of matter today, i.e., about $Y \approx 0.3$.

One can estimate the deuterium content of the matter from which the solar system formed by studying the amount of ^{3}He in the Sun. The point is that in the Sun, D transforms into ^{3}He upon capturing a proton. Thus, the amount of ^{3}He gives an upper limit on the amount of D in the presolar matter. Estimates give the value

$$\frac{\mathrm{D}}{\mathrm{H}} \approx (2.5 \pm 1.0) \times 10^{-5}. \tag{7.6.2}$$

On the other hand, the abundance of D in the atmosphere of Jupiter is estimated at D/H $\approx$ (2.9–7.5) $\times 10^{-5}$. Upper limits on the distribution of D in other celestial objects appear in Table 4.

The synthesis of deuterium proceeds on a large scale in main-sequence stars via the process $p + p \rightarrow \mathrm{D} + e^+ + \nu$, which is the first step of the hydrogen cycle in stars. However, the deuterium thus formed is immediately consumed through the reaction $p + \mathrm{D} \rightarrow {}^3\mathrm{He} + \gamma$. Both reactions have the same Coulomb barrier to overcome, and the difference in mass is not large, so the temperature dependence of both reactions differs little. The rate of the first reaction is proportional to $\exp[-15.6T_7^{-1/3}]$, and that of the second to $\exp[-7.2T_7^{-1/3}]$. But the first reaction proceeds via the weak interaction. Its rate is $\sim 3 \times 10^{16}$ times less than the rate of the second reaction. Hence, in a steady state (production = destruction), $n_\mathrm{D}/n_p \approx 3 \times 10^{-17}$. For typical values of the temperature ($T_7 \approx 1$) and density ($\rho \approx 100\ \mathrm{g\ cm^{-3}}$), the lifetime against depletion $[n_\mathrm{D}/(dn_\mathrm{D}/dt)]$ is very short: $\sim \frac{1}{6}$ s. Even for a temperature as low as 7×10^5 K and a density of 1 g cm^{-3}, the deuterium lifetime is only 10^6 yr. By a wide margin, therefore, the burning of D into ^{3}He predominates over the formation of D.

It is ordinarily considered that appreciable quantities of D cannot be synthesized in any galactic processes whatever or, if such things exist, in supermassive objects. Deuterium can apparently only be destroyed. Therefore, the synthesis of all the deuterium must have taken place at an early stage of the expansion of the hot Universe.

However, we note that Hoyle and Fowler (1973) recently suggested that D can be synthesized via the motion of semirelativistic α-particles through ionized hydrogen. Such conditions may be present in supernovae or in the explosions of supermassive stars. The authors come to the conclusion that all of the deuterium could in principle be synthesized in such processes.

Let us say a few words about other light elements. In theory, ^{3}He can

nthesized in the Galaxy by stellar processes. In contrast, ^{6}Li, ^{9}Be, ^{10}B, and also, probably, ^{11}B and a substantial part of ^{7}Li can be explained by processes involving the interaction of cosmic rays with matter.

Thus, it is likely that only the present abundance of ^{4}He and, to some lesser extent, that of D, can yield evidence of the primordial abundances. In other words, only ^{4}He and D are appropriate for a comparison of the theory of element synthesis in the hot Universe with observations made today. The respective abundances by mass of ^{4}He and of D at present (see equations (7.6.1) and (7.6.2)) are

$$Y \approx 0.3\,, \tag{7.6.3}$$

$$X_{\mathrm{D}} \approx 3 \times 10^{-5}\,. \tag{7.6.4}$$

What can one infer from these conclusions?

First of all, the observed value of Y is very close to the prediction of the theory in the simplest, most natural case, when there are no unknown particles and no noticeable leptonic charge (see the preceding section). It is necessary to emphasize that Y is not very sensitive to variations in the present day value of the average density of the Universe (see Fig. 32). This concurrence of theory and observation serves as a weighty argument in favor of the correctness of the theory.

The value $X_{\mathrm{D}} \approx 3 \times 10^{-5}$ can be obtained by the cosmological synthesis of deuterium only if the present-day density ρ_0 has a value $\sim 3 \times 10^{-31}$ g cm^{-3}, close to the minimum allowable value of the density due to the material composing the galaxies. If this is so, then intergalactic gas with a noticeable density cannot exist. (A noticeable density means one substantially greater than the density of the matter in galaxies—sufficient, for example, to close the Universe.) However, it is necessary to remember that the value of X_{D} for the prestellar matter is not known very reliably.

Supposing that the observed deuterium has a cosmological origin, Reeves *et al.* draw important conclusions from the fact that the deuterium abundance depends strongly on the matter density. Thus, the abundance by mass of D is

2×10^{-3} for $\rho_m = 5 \times 10^{-32}$ g cm^{-3}
$(\Omega_0 = 0.01\,;\quad H_0 = 50$ km s^{-1} Mpc$^{-1})$;

3×10^{-5} for $\rho_m = 5 \times 10^{-31}$ g cm^{-3}
$(\Omega_0 = 0.1\,;\quad H_0 = 50$ km s^{-1} Mpc$^{-1})$;

$\sim 10^{-8}$ for $\rho_m = 5 \times 10^{-30}$ g cm^{-3}
$(\Omega_0 = 1.0\,;\quad H_0 = 50$ km s^{-1} Mpc$^{-1})$.

All of the values are for the simplest model, the one with $n_\nu = n_{\bar{\nu}}$. The decrease in the deuterium density with increasing ρ_m is connected with the more rapid rate of deuterium burning under cosmological conditions.

The observed deuterium abundance argues in favor of $\Omega_0 \approx 0.1$. At the present time, however, the reliability of this result is not high. Nevertheless, this direction of research is extraordinarily important and promising. In the article by Hoyle and Fowler (1973), the question is posed of whether a part of the deuterium—or all of the deuterium—can be formed in explosions of supernovae and/or in shock waves propagating in the interstellar gas. In this study there is a weak point, though. The structure of the shock wave is considered in an oversimplified way; no account is taken of the plasma effects, and no other nuclear reactions are examined.

The discussion is not over. The present authors hope that the theory of cosmological deuterium formation will be put on a firmer basis and made more precise.

Hence, one can say that the theory of nucleosynthesis in cosmology is confirmed by observations, although the latter are still too rough to use as a basis for drawing more refined conclusions about the parameters of the cosmological model.

THE RADIATION DOMINATED PLASMA AND THE RELIC RADIATION

1 Introduction and General Review

We now set out to describe the last stages in the evolution of the Universe—those closest to the present.

In the first, radiation-dominated (RD) stage, space is filled with photons and a small admixture of electrons, protons, and helium nuclei. This stage begins with the annihilation of the positrons and concludes with the recombination of hydrogen: $e^- + p \rightarrow \mathrm{H} + \gamma$.

During the second stage, space is filled with photons and the neutral atoms H I and He I. In accord with the general plan of the book, we consider here the evolution of an idealized homogeneous model of the Universe, postponing until the next part an examination of perturbations away from homogeneity.

To what extent does such an examination correspond to reality? The answer to this question depends both on the assumptions about the perturbations and on the moment of time considered. At the present time it is clear that the perturbations are by no means small, and that the physical state of matter has nothing in common with the predictions of a strictly homogeneous model.

On the other hand, the mean free path of the relic-radiation (RR) photons is large; therefore, with high accuracy, RR remains homogeneous and isotropic (as observations show) even in the presence of strong inhomogeneities in the distribution of ordinary matter. Thus the effect of perturbations on RR can be partially examined within the realm of a homogeneous model.

One can consider perturbations—for example, a sprinkling of antimatter or the superposition of peculiar motions on the background of general expansion—as sources of energy. For energy is released by annihilation, while the kinetic energy of organized motion is transformed into heat.

These energy sources are inhomogeneously distributed in space. By virtue of the rapid mixing of photons, however, this inhomogeneity is not substantial. Hence, one can examine the change in the RR spectrum just as if the energy were released uniformly in a homogeneous model. This approximation also will be considered below. Thus, we will examine the period

This chapter was written jointly with R. A. Sunyaev.

$t > 10^3$ s, or $z < 10^8$, by which time nuclear reactions and the annihilation of positrons cease.

At first glance, this period is not interesting because we are considering a homogeneous Universe. We must, however, ask the reader to be patient and not fall asleep through the next passage, after which several up to date and interesting questions are considered.

In a completely homogeneous Universe, this period is characterized by a monotonic fall in the temperature from 3×10^8 K to the present day value of 2.7 K. These temperatures refer to the radiation whose spectrum at all times is of the equilibrium, Planck form. In the early stages, the nuclei and electrons are entirely disconnected from each other, but later, because of the cooling, recombination occurs. In the first approximation, this process is determined by the condition of thermodynamic equilibrium.

Instead of taking time as an independent variable, we prefer to use the corresponding redshift z, which is connected with the temperature by the formula $T = (1 + z) \times 2.7$ K. The connection between z and the time t depends on the magnitude of the matter density and on the Hubble parameter (see chap. 1).

For $T \approx 16{,}000$ K, the composition is as follows: the helium is 50% He II and 50% He III while the hydrogen is 100% H II. For $T \approx 7000$ K, the helium is 50% He I, 50% He II, and 0% He III, while the hydrogen is still 100% H II. The most interesting moment is when $T \approx 4000$ K, at $z \approx 1500$. At this time, the helium is 100% He I, but the hydrogen is 50% H I and 50% H II.

All of these calculations have been carried out according to the well known Saha formula (see § 8.2). This formula predicts the rapid disappearance of H II and of free electrons. In equilibrium, when $T \approx 3000$ K and $z \approx 1100$, there remains 0.1% of H II and free electrons; 99.9% of the protons having become H I (and, of course, all of the helium is He I).

The behavior of a neutral gas differs markedly from that of an ionized gas. For example, the interaction between matter and radiation is most easily maintained by free electrons. Without such electrons, a neutral gas hardly interacts with the radiation at all.

Upon expansion, a neutral gas cools adiabatically—with an adiabatic index of 5/3 (as opposed to 4/3 characteristic of radiation)—so that for our era ($z = 0$) a simple calculation would give $T_{\text{matter}} = 0.002$ K, a result which clearly does not conform to reality. Second, though the gas in galaxies and in clusters of galaxies in part consists of cold clouds of temperature $T \lesssim 100$ K, there is certainly no neutral hydrogen in the space between the clusters.

Recombination is very important to the growth of perturbations and to the growth of peculiar motions of matter. Before recombination, radiation strongly inhibits these processes. But, since we shall take up these questions thoroughly in the following chapters, we shall not dwell on them here.

The radiation spectrum has an equilibrium form at the present time not much because the processes leading to equilibrium are strong, as before, in a homogeneous Universe, there are no processes that cause deviations from equilibrium. There is only one exception, namely, the very process of hydrogen recombination; it is discussed in § 8.2.

Detailed calculations show that recombination lags behind equilibrium. The first evident consequence is the nonzero fraction of protons and electrons, of the order of 10^{-2} to $10^{-3}\%$ at low temperatures, whereas the equilibrium fraction would be negligibly small. The situation is similar to that encountered in the analysis of the survival of primordial quarks or antiquarks (see the preceding chapter). Less easily observable is the distortion, due to recombination, in the high-frequency part ($h\nu/kT > 30$) of the radiation spectrum. At the present time such high frequencies correspond to $\nu \gtrsim 1.5 \times 10^{12}$ Hz, or $\lambda \lesssim 0.02$ cm; at these frequencies the background radiation intensity is several times larger than that which corresponds to a Planck spectrum with $T = 2.7$ K.

This is unexpected because the total density of protons and electrons is some 10^8 times smaller than the density of photons. One can, however, reconcile oneself to this result if one keeps in mind that the relatively large increase in the radiation intensity occurs only in a region of the spectrum which contains a very small fraction of all the photons (10^{-10} of all of the photons have frequencies satisfying $h\nu/kT > 30$). There is little hope that this prediction of the theory will be tested by observations because there are many sources of background radiation with $h\nu/kT \gg 1$, primarily galactic nuclei and dust. The radiation from the latter is most strong in the spectral region $\lambda < 3 \times 10^{-2}$ cm. Many quasars radiate in this same spectral region. Finally, there are still other processes leading to a distortion of the high-frequency region of the spectrum.

Now the theory of a completely homogeneous Universe leads to a picture of cold, homogeneously distributed atomic hydrogen immersed in a background of blackbody radiation. This dull picture is very far from the observed Universe of stars, galaxies, and quasars; from the active processes of nuclear burning and gravitational interaction; and from the large-scale motions and outbursts that lead to the emission of radio waves, visible light, X-rays, and cosmic rays.

We understand that all of this activity is a result of the original inhomogeneity, which leads to the birth of astronomical objects. Additionally one can think of other types of energy-producing perturbations occurring in the early stages of the evolution. For example: (1) an inhomogeneous distribution of antibaryons would lead to a delay in the annihilation; (2) inhomogeneities in the density distribution would give rise to sound and shock waves. In the literature, vortex (turbulent) motion is also considered. Finally, there exist homogeneous, anisotropic cosmological models whose

transformation into a homogeneous model is accompanied by the liberation of energy—more precisely, by the growth in entropy.

It was noted earlier that, with respect to the influences on the radiation spectrum, many of the problems studied can be considered within the context of the homogeneous theory. This is true even if the energy sources arise as a result of inhomogeneity. Indeed, when energy (in the form of radiation, cosmic rays, or shock waves) is scattered in the surrounding space, it becomes homogeneously distributed in a comparatively short time.

Some sources of energy at first give the energy to free electrons—directly or after the ionization of neutral atoms. The entire subsequent discussion is based on the interaction of this plasma with the radiation. Since we are concerned with a very rarefied plasma, the radiation and absorption of photons as a consequence of $p + e^-$ and $\alpha + e^-$ collisions is less important than that due to the Compton effect—the scattering of photons by electrons. Hence, § 8.2 is devoted to an account of the general theory of the velocity distribution of electrons in an arbitrary radiation field. It is shown there that the electrons have a Maxwellian distribution characterized by a definite temperature. Then, in § 8.4, we will examine the kinetic equation for photons, remembering always that the scattering occurs from a Maxwellian distribution of electrons. This theory was developed mainly on the basis of the work of A. S. Kompaneets (1956) and Weymann (1965).

Passing to cosmology, we can expect three different situations:

1) Energy is liberated very early ($z > 10^6$) when the electron density is sufficiently high to establish new equilibria, indistinguishable from the initial equilibrium. Such a situation occurs as a result of e^+e^- annihilation; the entropy of the radiation grows ($S_{\gamma 0\ \text{after annih}} = [S_\gamma + S_{e^+} + S_{e^-}]_{\text{before annih}}$), but the spectrum remains Planckian.

2) Energy is liberated in the interval $10^4 < z < 10^6$. The contribution of scattering is sufficiently large to create equilibria, but new photons are created insufficiently fast. Hence the temperature of the electrons is always much greater than the radiation temperature. Clearly, the scattering redistributes the photons over the spectrum, but does not change their density. Consequently, only a partial equilibrium is achieved here for a given density of photons. This situation leads (see § 8.5) to a Bose-Einstein spectrum:

$$F_\nu^{\text{BE}} = \frac{2h\nu^3}{c^2}\left[\exp\left(\frac{h\nu + \mu}{kT}\right) - 1\right]^{-1} \tag{8.1.1}$$

(notice that $\mu > 0$ in the exponent!), instead of a genuine Planck spectrum:

$$F_\nu^{P} = \frac{2h\nu^3}{c^2}\left[\exp\left(\frac{h\nu}{kT}\right) - 1\right]^{-1}. \tag{8.1.2}$$

The deviations from the Planck formula are strongest in the Rayleigh-Jeans region ($h\nu < kT$). This is the best studied region of the RR spectrum;

distortions have not been observed. Such a null result gives an upper limit to the magnitude of the early liberation of energy.

If energy is liberated for $z < 1500$, i.e., after recombination. The gas then rapidly ionizes a second time. For a hydrogen plasma to be ionized by means of electron impacts, an electron temperature higher than 10^4 K is required, a temperature much greater than that of RR at that time. As a consequence of the low electron density and of the small amount of scattering, equilibrium cannot be established again. The kinetic equation for photons predicts definite distortions in the spectrum, mainly for $h\nu \gtrsim 2kT$. Experimental data in this region of the spectrum were for a long time contradictory. There is presently no indication of distortion, but the accuracy is still small. These questions are discussed in § 5.3.

The general conclusions of §§ 8.4–8.6 are then as follows: Soon after the discovery of RR, it was believed that the radiation must be precisely Planckian. Now it has become clear that deviations from this spectrum (not yet found, however) would not contradict the predictions of the hot model of the Universe. Indeed, they can give valuable additional information.

The physical conditions of the intracluster gas (the gas inside of clusters of galaxies) are of evident interest, for research on the radiation from this gas is relevant to estimates of the density of diffuse matter in the Universe. This matter is examined in § 8.7.

It is well known that the density of diffuse gas (H I) is negligibly small. For distant quasars ($z > 2$), the Lyman-α line falls in the spectral region $\lambda \geq (1 + z)\,\lambda_0 > 3600$ Å which passes freely through the Earth's atmosphere. Thus, the study of quasar spectra informs us of the absence of neutral atomic hydrogen—its density turns out to be $n_H < 10^{-11}$ cm^{-3}. With less accuracy, the same conclusions follow from radio-astronomical observations of the 21 cm line of H I.

In the theoretical work of Ginzburg and Ozernoy (1965), it is shown that cosmic rays can ionize and heat the intergalactic gas. The loss of energy by the gas and the recombination rate under present-day conditions ($\rho < 10^{-29}$ g cm^{-3}) are so small that total ionization is very probable.

The radiation from the completely ionized hydrogen (and possibly also from helium) constituting the intergalactic gas is the subject of many studies. There are two unknown parameters—the density of the gas, and its temperature. At high temperatures the radiation is dominated by X-rays, while at much lower temperatures it is dominated by ultraviolet emission lines. A continuous distribution of gas in space gives, when account is taken of redshift, emission lines forming a wide band. Finally, such a gas also produces continuous radio-wave radiation having a flat spectrum that differs sharply from a Planck spectrum.

A detailed analysis of the observations cited by Sunyaev in 1969*b* leads to the result that the density of gas is less than (0.2–0.3) ρ_c ($\rho_c = 3H_0^2/8\pi G =$

$2 \times 10^{-29} h_0^2$ g cm^{-3}). But for the sake of generality we will use in the subsequent development the parameter $\Omega_g = \rho_g/\rho_c$, the dimensionless, present-day gas density.

All of these conclusions are based on the measurements by Kurt and Sunyaev (1967*a*, *b*), using Soviet space probes, of the ultraviolet background, and on the analysis by Sunyaev of the rarefied hydrogen halo of the Galaxy (which serves as a very sensitive detector of hard ultraviolet and soft X-ray radiation). The principal conclusion drawn from this series of studies is that the Universe is probably open, i.e., hyperbolic (see chap. 2).

During the last four years, the original estimates of the ionizing radiation falling on the Galaxy have continued to hold. It is possible that by observing the halo more carefully in a region of low H I density, we will even manage to lower the upper limit on the radiation flux. However, a final conclusion concerning the open character of the Universe is today open to question because of the revision in the value of the Hubble constant. In Sunyaev's original work (1969*a*), the calculations were carried out using $H_0 = 100$ km s^{-1} Mpc^{-1}. If one takes the other extreme value found in the literature today, $H_0 = 50$ km s^{-1} Mpc^{-1}, then the critical density is four times smaller ($\rho_c \sim H_0^2$). Therefore, inferences drawn from an examination of the halo of the Galaxy do not exclude the flat model of the Universe in which the basic matter is found in the form of completely ionized, rarefied gas more or less uniformly distributed between clusters of galaxies.

In § 8.8 the X-rays emitted by a rarefied gas are studied. The results of the calculation depend, naturally, on assumptions about the gas density, for which only an upper limit is known. The general conclusion reached is that the hard X-ray background observed today significantly exceeds the possible radiation from the gas.

The isotropic X-ray background apparently represents the radiation from a large number of discrete sources (Longair and Sunyaev 1969*a*, *b*). An increase in the angular resolution and in the sensitivity of the detectors will finally permit one to solve this problem.

Finally, we examine in § 8.9 a question that stands off by itself—the question of the interaction of cosmic rays with the radiation background. This background consists first and foremost of RR.

The loss of energy by cosmic rays limits the maximum energy of the cosmic ray particles; this limit depends on the age of the cosmic rays. In the period from 1969 through 1971, rocket experiments overestimated the total density of RR by a factor of 20 to 100. Indeed, results from ground-based observations of cosmic rays with energies up to 10^{22} eV were known, in evident contradiction with the results of the rocket experiments. These contradictions were largely eliminated after experiments done by the Los Alamos group (Blair *et al.* 1971).

Another conclusion following from the examination of cosmic-ray interactions is that cosmic rays are a young constituent of the Universe—they

cannot have a primordial origin. For the strong energy exchange between charged particles and photons in the past (see chap. 6) excludes the possibility of large energies for charged particles in the early stages of the radiation dominated era before recombination.

8.2 Recombination: Equilibrium and Kinetics

Let us first write Saha's equilibrium formula:

$$\frac{n_e n_p}{n_H} = \frac{g_e g_p}{g_H} \frac{(2\pi m_e kT)^{3/2}}{h^3} e^{-I/kT} \tag{8.2.1}$$

for the reaction $e^- + p \rightleftarrows H + \gamma$. The ionization energy I is 13.6 eV, so that $I/k = 1.58 \times 10^5$ K; the statistical weight factor is $g_e g_p / g_H = 1$. Finally, substituting the numerical values, we obtain

$$\frac{n_e n_p}{n_H} = 2.4 \times 10^{15} T^{3/2} \exp(-1.58 \times 10^5/T) . \tag{8.2.2}$$

Now let us introduce the degree of ionization, $x \equiv n_p/(n_p + n_H)$. The density is expressed in terms of the temperature through the formula $n \sim T^3$; therefore, at the present time, $n_0 \approx 10^{-5}\Omega_0 h_0^2$ cm^{-3}, $T = 2.7$ K, and consequently, $n \approx 6 \times 10^{-7}(\Omega_0 h_0^2) T^3$. This leads to the final form

$$\frac{x^2}{1-x} \approx 4 \times 10^{21}(\Omega_0 h_0^2)^{-1} T^{-3/2} \exp(-1.58 \times 10^5/T) . \tag{8.2.3}$$

The equilibrium values of $x(T)$ are given in Table 5 for $\Omega_0 h_0^2 = 1$ and $\Omega_0 h_0^2 = 1/40$.

The process of recombination, which we will now study, has been examined by Zel'dovich, Kurt, and Sunyaev (1968); by Peebles (1968); and by Sunyaev and Zel'dovich (1970*a*).

The photoionization process, $H + \gamma \rightarrow p + e^-$, and the reverse process of radiative recombination, $p + e^- \rightarrow H + \gamma$, play a basic role in the maintenance of the equilibrium of an ionized gas under cosmological conditions. The Saha formula naturally does not depend on the specific mecha-

TABLE 5

Degree of Ionization x of Atomic Hydrogen for Various Equilibrium Conditions

Temperature (K)	$x(\Omega_0 h_0^2 = 1)$	$x(\Omega_0 h_0^2 = 1/40)$
5000	0.996	0.9999
4500	0.9	0.997
4000	0.3	0.85
3500	0.027	0.17
3000	0.0007	0.004
2500	10^{-6}	8×10^{-6}

nism for the establishment of equilibrium. It has the same form in a dense gas, where collisional ionization and three-body recombination via the collisions $H + e^- \rightleftarrows p + e^- + e^-$ occur, as it does in a diffuse gas. However, in order for the Saha formula to be satisfied under cosmological conditions, it is necessary that the radiation spectrum be of the Planck form.

The kinetics describing the establishment of equilibrium and the approach of the actual densities to their equilibrium values is determined by two factors: (1) the rate of ionization and recombination, and (2) the deviations of the radiation spectrum from the Planck form. Calculation shows that for a Planck radiation spectrum, the equilibrium corresponding to the values in Table 5 persists down to $x \approx 3 \times 10^{-4}$ for $\Omega_0 h_0^2 = 1$ and to $x \approx 10^{-2}$ for $\Omega_0 h_0^2 = 1/40$.

The condition for the maintenance of equilibrium has the form

$$\omega\tau = \alpha_{\text{tot}} x n t (3kT/I) . \tag{8.2.4}$$

Here ω is the recombination probability per unit time and τ is the e-folding time. The former is $\omega = \alpha_{\text{tot}} n_{e^-} = \alpha_{\text{tot}} x n$, where α_{tot} is the recombination rate coefficient averaged over a Maxwellian distribution, $\alpha_{\text{tot}} \approx 4 \times 10^{-11} T^{-1/2}\ \text{cm}^3\ \text{s}^{-1}$ (see Kaplan and Pikel'ner 1963). The e-folding time satisfies $\tau = (3kT/I)t$, with t the cosmological time, because $x^2 \sim e^{-I/kT}$ and $T \sim z \sim t^{-2/3}$, so that

$$\tau^{-1} = -\frac{d \ln x}{dt} = -\frac{1}{2}\frac{I}{kT}\frac{d \ln T}{dt} = \frac{1}{3}\frac{I}{kT}\frac{1}{t} .$$

Still earlier, however, the process of recombination itself changes the radiation spectrum. And the part of the spectrum (the short-wave section) most affected is the region containing those photons that participate in the reverse process of photoionization. This last process then influences the densities n_{e^-} and n_p. The total number of photons is very large; consequently, the spectral changes can affect only a small fraction of them. But for direct recombination to the ground state ($1S$) level, photons are actually emitted with an energy I. Since recombination occurs for $kT \ll I$, the number of such photons in equilibrium, proportional to $\exp(-I/kT)$, would be negligibly small, much less than the number of electrons and protons. The part of the spectrum important for the process actually changes substantially as a consequence of the recombination, if one assumes that recombining electrons proceed directly to the ground state. In turn, this spectral change causes a shift (an increase) in the density of electrons. Cascading recombination in principle moderates the deviations if, instead of one "high-energy" photon, many of "low energy" are obtained. However, in the case of hydrogen, radiative recombination inevitably gives a single photon with energy less than that of a Lyman-α photon, i.e., with an energy less than $\frac{3}{4}I$ for each single atom formed in the ground state. Therefore, the recombination retardation effect remains.

The kinetics is determined by the case with which recombination to an excited level of atomic hydrogen occurs. A similar statement holds for the ionization of excited atoms. The equilibria $\gamma + H^* \rightleftharpoons p + e^-$ and $\gamma + H^* \rightleftharpoons H^{**}$ are thus maintained over a long time period. For these processes, photons with energies $I/4 = 3.4$ eV or less are necessary, and there are a sufficient number of collisions with these in the period of time considered. But liberation of the excitation energy, corresponding to the Lyman α line ($\frac{3}{4}I \approx 10.2$ eV), or of even more energy, is a difficult step.

If recombination occurs directly to the ground state, the photons emitted have energies $E = I$. Such photons rapidly ionize other neutral atoms. Even in the cascade processes

$$e^- + p \rightarrow H^* + \gamma_1 , \qquad H^* \rightarrow H(1S) + \gamma_2 ,$$

the energy $E_{\gamma,2}$ satisfies $E_{\gamma,2} \geq \frac{3}{4}I$, so that the photon γ_2 can be absorbed by an atom in the ground state, thereupon forming excited hydrogen, H^*, which is then easily ionized by the fundamental part of the radiation.

To examine this further, we introduce β, the sum of the densities of free protons, excited atoms, and energetic photons (we call a photon energetic if its energy is greater than or equal to that of a Lyman-α photon, i.e., $0.75I$) normalized with respect to the total density

$$\beta = \frac{n_p + n_{H^*} + n_{\gamma^*}}{n_p + n_{H^*} + n_{H(1S)}} = \frac{n_p + n_{H^*} + n_{\gamma^*}}{n} . \tag{8.2.5}$$

Now thermodynamic equilibrium is maintained among the various constituents, so with $n_p = n_{e^-}$,

$$\begin{aligned} \frac{n_p n_{e^-}}{n_{H(2P)}} &= \tfrac{1}{3} \times 2.4 \times 10^{15} T^{3/2} \exp(-39400/T) , \\ \frac{n_{H(nl)}}{n_{H(2P)}} &= \frac{g_{nl}}{g_{2P}} \exp[-(E_n - E_{2P})/kT] . \end{aligned} \tag{8.2.6}$$

The density n_{γ^*} is given roughly by an integral over the Planck function:

$$n_{\gamma^*} = \int_{\nu_2}^{\infty} \frac{8\pi\nu^2}{c^3} [\exp(h\nu/kT_{\text{eff}}) - 1]^{-1} \, d\nu , \tag{8.2.7}$$

where T_{eff} is determined by the formula

$$\frac{n_{H(2P)}}{n_{H(1S)}} = \frac{g_{2P}}{g_{1S}} \exp\left[-\frac{(E_{2P} - E_{1S})}{kT_{\text{eff}}}\right] = 3 \exp\left[\frac{-118{,}000 \text{ K}}{T_{\text{eff}}}\right] . \tag{8.2.8}$$

Note that by excluding T_{eff} from the formulae, we could obtain an expression similar to the mass-action law $n_p n_{e^-}/n_{H(1S)} n_{\gamma^*} = f_1(T)$ or to the one $n_{H(2P)}/n_{H(1S)} n_{\gamma^*} = f_2(T)$. In the early stages, T_{eff} is several times larger than T.

Now it is important that free protons are the most abundant particles entering into the expression for β. For example, if we specify complete

equilibrium, $T_{\rm eff} = T$, and take the values $\Omega_0 h_0^2 = 1$ and $x = 0.1$, we obtain the following densities:

$$n_{\rm H(1S)} = 2.5 \times 10^4\ {\rm cm}^{-3}, \quad n_p = n_{e^-} = 2.5 \times 10^3\ {\rm cm}^{-3},$$

$$n_{\gamma,\rm tot} = 10^{12}\ {\rm cm}^{-3}, \quad n_{\gamma^*} = 10\ {\rm cm}^{-3},$$

$$n_{\rm H(2P)} = 2 \times 10^{-9}\ {\rm cm}^{-3}, \quad n_{\rm H^*} = \sum_{n=2}^{n_m} n_{{\rm H},n} = 3 \times 10^{-7}\ {\rm cm}^{-3}.$$

In calculating $n_{\rm H^*}$, we found an upper limit n_m on the principal quantum number according to the Ingles-Taylor formula $n_m \approx 10^3\ (n_{\rm H})^{-2/15}$. This takes into account the merging of levels that is a consequence of the Stark effect in hydrogen.

If formula (8.2.7) is considered as a definition of $T_{\rm eff}$, then equation (8.2.8) is the physical statement that the ratio $n_{\rm H(2P)}/n_{\rm H(1S)}$ depends on the number of energetic photons, but not on the whole spectrum of photons. Notice that $n_{\gamma,\rm tot} \gg n_p, n_{e^-}, n_{\rm H(1S)}$, but $n_{\gamma^*} \ll n_p, n_{e^-}, n_{\rm H(1S)}$. Consequently, one can identify the degree of ionization x with the variable β, in the subsequent development.

In the respective kinetic equations for $n_p/n_{\rm H}$, $n_{\rm H^*}/n_{\rm H}$, and $n_{\gamma^*}/n_{\rm H}$, there is a large number of terms representing rapid processes, as, for example, $e^- + p \rightleftarrows {\rm H}^* + \gamma$ and ${\rm H}_{(2P)} \rightleftarrows H_{(1S)} + \gamma^*$. These terms cancel when we pass to the equation for the sum β; indeed this is just the reason for the introduction of β. This sum decreases as a result of two processes only: (1) the redshift due to the expansion of the Universe, which transforms energetic photons γ^* into photons that cannot excite normal hydrogen; and (2) the two-photon decay of the metastable $2S$ state of hydrogen (Kipper 1950; Spitzer and Greenstein 1951). The rate of this latter process is $\omega_{2S} \approx 8\ {\rm s}^{-1}$.

Consequently,

$$\frac{d\beta}{dt} = -\left|\frac{d\nu}{dt}\right| \frac{n_{\gamma^*}}{\Delta\nu n} - \omega_{2S} \frac{(n_{\rm H(2S)} - n_{\rm H(2S)eq})}{n}. \tag{8.2.9}$$

In this equation, $n_{\rm H(2S)}$ and n_{γ^*} are expressed through β using the equilibrium condition: $n_{\rm H(2S)eq} = n_{\rm H(1S)} \exp\ (-118{,}000/T) = n_{\rm H(2P)}$. The introduction of $\omega_{2S} n_{\rm H(2S)eq}$ on the right-hand side of the equation permits a description of the reverse process: ${\rm H}_{(1S)} + 2\gamma \rightarrow {\rm H}_{(2S)}$. The first term, describing the redshift, is relatively small. Finally, we obtain

$$\begin{aligned} \frac{dx}{dt} &= -\omega_{2S} \frac{(2\pi h)^3}{(2\pi m_e kT)^{3/2}} n e^{I/4kt}(x^2 - x_{\rm eq}^2) \\ &= 2 \times 10^{-21}(\Omega_0 h_0^2)(T/\ {\rm K})^{3/2} e^{39{,}400/T}(x - x_{\rm eq})^2. \end{aligned} \tag{8.2.10}$$

The difference between x and $x_{\rm eq}$ is small at high temperatures $T > 4000$ K where $x > 0.5$. Later, $x \gg x_{\rm eq}$ so that one can neglect $x_{\rm eq}$ in the paren-

theories. In this later period, the approximate formula (Sunyaev and Zel'dovich 1970*a*; see also Longair and Sunyaev 1969):

$$x = \frac{1.6 \times 10^7}{(\Omega_0 h_0^2)^{1/2} T} e^{-A/T}$$

holds, where $A = I/4k = 39{,}400$ K. Finally, for $T < 2400$ K, there is no longer equilibrium among p, e^-, and H(2S). The results are collected in Table 6 for $\Omega_0 h_0^2 = 1$ (compare these with x_{eq} of Table 5).

The calculation gives $x \to 5 \times 10^{-5}$ for $t \to \infty$ and $T \to 0$. For low temperatures, i.e., when $T < 200$ K, the process $H\,I + e^- \to H^- + \gamma$ and, further, $H^- + H \to H_2 + e^-$ are possible thermodynamically. For even still lower temperatures ($T < 100$ K) the molecule H_2 is thermodynamically stable. But the formation rate of H^- and H_2 is so low that in a low-density homogeneous Universe they are essentially not formed. This question will arise again later in connection with the problem of galaxy formation.

The recombination of hydrogen leads to a surplus of short-wave photons with respect to the number expected on the basis of the Planck spectrum. In Table 7 several points from the expected spectrum are given (for more detail, see the original works cited above). The effect is relatively large in regions where the absolute fluxes, i.e., both the equilibrium flux $F_{\nu,eq}$ and that calculated taking account of the recombination, $F_{\nu,cal}$, are small. The wavelengths and fluxes of the radiation are in the units microns and ergs $(cm^2\ s\ Hz)^{-1}$, respectively, and refer to the present time.

TABLE 6

DEGREE OF IONIZATION x FOR NON-EQUILIBRIUM CONDITIONS

Temperature (K)	x
5000	0.996
4500	0.924
4000	0.40
3500	0.07
3000	0.0098
2500	9×10^{-4}
2000	10^{-4}
1500	5×10^{-5}

TABLE 7

EFFECT OF THE RECOMBINATION ON THE BACKGROUND RADIATION

Wavelength (microns)	$h\nu/kT$	$F_{\nu,eq}$ (ergs cm^{-2} s^{-1} Hz^{-1})	$F_{\nu,calc}$ (ergs cm^{-2} s^{-1} Hz^{-1})
200	27	1×10^{-22}	2×10^{-22}
150	36	3×10^{-26}	5×10^{-24}

8.3 The Interaction between the Electrons and the Photons in the Dilute Plasma

In the ideal, strictly homogeneous model of the Universe, recombination takes place in the manner described in § 8.2, when $z \approx 1000$, $T \approx 3000$ K. Some residual ionization, $\alpha_\infty \approx 10^{-4}$ to 10^{-5}, remains (the exact value depends on Ω). Meanwhile, the temperature of the weakly ionized gas follows the radiation temperature down to $T \approx 400$ K when $z \approx 150$, this by virtue of the energy exchange mediated by Compton scattering. In the subsequent evolution, the thermal exchange markedly weakens and the gas cools adiabatically, $T \sim (1 + z)^2$. This results in $T_0 = 400/(150)^2$ K $= 2 \times 10^{-2}$ K today. But such an idealized description differs quite sharply from reality. A significant part of the gas actually transforms into stars, galaxies, quasars, and other such dense bodies. The remaining gas heats up and ionizes, and is then not in equilibrium with the relic radiation, the temperature of which is low. In this section, the process leading to the cooling of the plasma is examined. This plasma is embedded in a radiation field characterized by a lower temperature. There does, however, exist the energy necessary to maintain the plasma at a high temperature.

In particular, for each given plasma temperature, the corresponding quasi-static degree of ionization is calculated. The Saha formula does not apply in this case, *because* the ionization results from electron collisions while the recombination is accompanied by the emission of photons. The Elvert formula

$$\frac{n_{\mathrm{H\,II}}}{n_{\mathrm{H\,I}}} = 6T \cdot 10^{-69000/T} \tag{8.3.1}$$

therefore holds, with analogous expressions for the concentrations of He II and He III relative to that of He I. This formula gives almost complete ionization of the hydrogen for temperatures of 10^5 K and 10^6 K, but little ($\sim 10^{-2}$) ionization at 10^4 K, in distinction to the Saha formula prediction. An assumption concerning a 10^4 K temperature for the plasma therefore does not agree with the observations.

The main processes in a completely ionized, nonrelativistic plasma (4000 K $< T < 10^8$ K) which interacts with soft photons ($h\nu < 0.1$ MeV) are (1) bremsstrahlung, (2) Compton scattering, and (3) recombination radiation and photoionization.

The energy radiation rate of the bremsstrahlung processes $e^- + z \rightarrow e^- + z + \gamma$ is characterized by the differential cross section for the radiation of a photon in the interval between E and $E + dE$,

$$d\sigma = \frac{16\pi^2}{3^{3/2}} \frac{z^2 e^6 g(E)}{m_e c^3 E_0 E}\, dE\,. \tag{8.3.2}$$

Here $g(E) = 1$ for $E \gg 2^{1/2}E_0^{3/2}h/(\pi^2 z e^2 m^{1/2})$, $g(E) = 0$ for $E > E_0$, and

$$g(E) = \frac{\ln 2^{1/2}E_0^{3/2}h}{1.78\pi z e^2 m^{1/2}E} \quad \text{for} \quad E \ll \frac{2^{1/2}E_0^{3/2}h}{\pi z e^2}.$$

The initial energy of the electron is E_0, and z is the charge of the nucleus. The number of photons created per unit time is proportional to the integral $\int (d\sigma/dE)dE$, which diverges because of the radiation of a large number of soft photons. An examination of the energy balance, however, shows that the average cross section, determined by a different formula, actually enters here. This quantity,

$$\langle\sigma\rangle = \frac{1}{E_0}\int E\left(\frac{d\sigma}{dE}\right)dE = \frac{16\pi^2 z^2 e^6\langle g\rangle}{3^{3/2}m_e c^2 E_0^2}, \tag{8.3.3}$$

has great significance because it determines the energy loss rate of the electrons starting from a given initial energy E_0. For a Maxwellian distribution of electrons and singly charged ions (protons),

$$\frac{d\epsilon}{dt} = -n_e n_p\langle\sigma v\rangle E_0 = 1.43 \times 10^{-27}\langle g\rangle T^{1/2}n_e n_p \text{ ergs cm}^{-3}\text{ s}^{-1}, \tag{8.3.4}$$

where $\langle g\rangle \approx 0.8$. Let us also introduce the characteristic time for energy loss through bremsstrahlung (free-free scattering),

$$\tau_{ff} = \frac{3nkT}{(d\epsilon/dt)} = 3.6 \times 10^{11}T^{1/2}n_p^{-1}\text{ s}. \tag{8.3.5}$$

Above we considered only bremsstrahlung emission. Such a calculation applies only to an optically thin plasma, i.e., to conditions such that photons leave the plasma quickly. In the opposite case, when, in the volume considered, the radiation is in equilibrium, at a temperature T_r, and the electrons are cold, bremsstrahlung absorption dominates and the electrons are heated. It is evident that with equality between the electron and radiation temperatures, equilibrium holds and the exchange of energy ceases. Thus, with account of bremsstrahlung absorption, we have approximately

$$\frac{dT_e}{dt} = \frac{T_r - T_e}{\tau_{ff}}, \tag{8.3.6}$$

with T_e, not T_r, entering the expression for τ_{ff}.

Let us now consider another process essential in the presence of photons, namely, Compton scattering. This process has a scattering cross section (the Thomson cross section) independent of the frequency (in the nonrelativistic limit). It is

$$\sigma_c = \sigma_T = \frac{8\pi}{3}\left(\frac{e^2}{m_e c^2}\right)^2 = 6.65 \times 10^{-25}\text{ cm}^2. \tag{8.3.7}$$

The angular dependence of the differential cross sections is weak and symmetric (θ is the scattering angle and Ω is the solid angle):

$$d\sigma_c = \text{const.}\,(1 + \cos^2\theta)d\Omega. \tag{8.3.8}$$

Now let us examine the average force acting on an electron at rest in an anisotropic radiation field, i.e., in a field whose intensity depends on direction. The radiation propagating in the direction $\boldsymbol{n}$ carries in this direction an energy flux $\phi\boldsymbol{n}$ and a momentum flux $\phi\boldsymbol{n}/c$. The quantity ϕ has the dimensions ergs $\text{cm}^{-2}\,\text{s}^{-1}$.

Scattered radiation, by virtue of the symmetry of the scattering law, has vanishing momentum. The momentum given to an electron per unit of time, i.e., the force acting on an electron, is $\boldsymbol{f} = (\sigma_e/c)\int\phi(\boldsymbol{n})\boldsymbol{n}d\Omega$. Passing to the angular distribution of the radiation, we obtain [note that we are here changing the notation; under the integral, ϕ has the dimensions ergs $(\text{cm}^2\,\text{s sr})^{-1}$]

$$\boldsymbol{f} = (\sigma_e/c)\int\phi(\boldsymbol{n})\boldsymbol{n}d\Omega\,,$$

where the integration is carried out over solid angle.

One can compare this expression with the expression for the energy flux in an arbitrary radiation field,

$$\boldsymbol{q} = \int\phi(\boldsymbol{n})\boldsymbol{n}d\Omega\,, \tag{8.3.9}$$

to show that

$$\boldsymbol{f} = \boldsymbol{q}\sigma_e/c\,. \tag{8.3.10}$$

This expression was used in Volume 1 for an examination of light pressure, for an examination of the Eddington limit of luminosity, and for other problems connected with fluxes of radiant energy. Below it will be shown that the force also consists of a term proportional to the square of the radiation flux. Here we are considering the case of small $\boldsymbol{q}$, so the quadratic effect is negligible.

Let us now pose a somewhat different problem—one characteristic of cosmology. Suppose that isotropic equilibrium radiation, with no mean motion (with vanishing energy flux), fills space. In this space an electron moves with velocity $\boldsymbol{\mu}$, where $|\boldsymbol{\mu}| \ll c$. What is the force acting on this electron?

We reduce the problem to the previous one by passing to a coordinate system in which the electron is at rest. The energy flux in the Lorentz-transformed coordinate system is $\boldsymbol{q} = -4\epsilon_r\boldsymbol{\mu}/3$, where ϵ_r is the radiation energy density. The minus sign signifies that in the rest system of the electron, the flux is directed opposite to the direction of motion of the electron in the isotropic field. The factor 4/3 appears because the work done by pressure forces, $-\boldsymbol{\mu}P = -\boldsymbol{\mu}\epsilon_r/3$, is added to the energy flux. This factor 4/3 also appears when formally Lorentz-transforming the stress-energy tensor T_{ik} according to the rules of tensor analysis, if, at the same time, we neglect the square of the velocity μ^2/c^2. Hence, the average force and corresponding acceleration are

$$\boldsymbol{f} = m\frac{d\boldsymbol{\mu}}{dt} = \boldsymbol{q}\sigma_e/c = -\tfrac{4}{3}\sigma_e\epsilon\boldsymbol{\mu}\,. \tag{8.3.11}$$

One can now introduce the decay time for the velocity. It is more convenient, however, to write the equation for the kinetic energy of the electron after multiplying both sides of equation (8.3.11) by μ. Thus, electrons, with kinetic energy E, lose energy due to collisions with photons at the rate

$$\frac{dE}{dt} = -\mu m \frac{d\mu}{dt} = -\tfrac{8}{3} E \frac{\sigma_c \epsilon_r}{m_e c} . \tag{8.3.12}$$

Here ϵ_r is the energy density of the radiation field.

The characteristic time for energy loss due to Compton scattering is

$$\tau_c = \frac{3 m_e c}{8 \sigma_c \epsilon} . \tag{8.3.13}$$

But the energy-decrease law written above is not exact: the energy of an electron falls not to zero but to that value of the thermal energy corresponding to equality between the temperatures of the electrons and the radiation. Consequently, the exact energy-decrease equation has the form

$$\frac{dE}{dt} = (\tfrac{3}{2} k T_r - E)/\tau_c . \tag{8.3.14}$$

The physical reason for this is that in a radiation field, an electron feels a chaotic force due to random scattering with photons as well as an average braking force. The chaotic force increases the electron energy. The simple formula (8.3.14) then applies to an electron in the field of Planck equilibrium radiation.

The role of the third process, recombination and line emission, is important at low temperatures. For $T = 10^4$ K, the radiation in the recombination continuum and lines is approximately 100 times stronger than the energy loss due to the slowing-down process.

Finally, the time for the exchange of energy between electrons is given by the formula

$$\tau_{ee} = (\sigma_{ee} n_e v)^{-1} = n_e^{-1} \left(\frac{m}{kT_e}\right)^{1/2} \left(\frac{e^2}{kT_e}\right)^{-2} (\pi \ln \Lambda)^{-1} . \tag{8.3.15}$$

The factor $\ln \Lambda$ takes account of the effects of distant collisions (see Spitzer 1962) and is of the order 30.

Order-of-magnitude parameters for these various mechanisms are given in Table 8.

As an example, let us compare τ_{ff}, τ_c, and τ_{ee} in cases characteristic of cosmology; we choose $\Omega_0 = 1$ and $z = 0$, 30, and 1000. Note that the subscript "ff" refers to free-free or bremsstrahlung. The matter considered consists of 30% He by mass. Moreover, τ_c does not depend on Ω_0 whereas τ_{ff} and τ_{ee} are inversely proportional to Ω_0. The results in Table 8 show that electron collisions are the most rapid process. The energy exchange between electrons leads to a Maxwellian distribution of their energy. But these collisions do not change the total energy of the electrons, and con-

TABLE 8

CHARACTERISTIC TIMES FOR VARIOUS ENERGY-LOSS MECHANISMS IN A RADIATION-FILLED PLASMA

z	ϵ (ergs cm^{-3})	n_e (cm^{-3})	t[a] (s)	τ_c (s)	τ_{ff} (s)	$\tau_{fb}(\tau_{fb}[H])$[b] (s)	$\tau_{bb}(\tau_{bb}[H])$[c] (s)	τ_{ee} (s)
					$T=10^5$ K			
0	4×10^{-13}	6×10^{-6}	4×10^{17}	4×10^{19}	2×10^{19}	$2\times10^{18}(6\times10^{18})$	$4\times10^{17}(7\times10^{18})$	5×10^{10}
30	4×10^{-7}	0.2	2.4×10^{15}	4×10^{13}	6×10^{14}	$7\times10^{13}(2\times10^{14})$	$1\times10^{13}(2\times10^{14})$	1.5×10^{6}
1000	0.4	6×10^{3}	1.3×10^{13}	4×10^{7}	2×10^{10}	$2\times10^{9}(6\times10^{9})$	$4\times10^{8}(7\times10^{9})$	50
					$T=10^6$ K			
0	4×10^{-13}	6×10^{-6}	4×10^{17}	4×10^{19}	6×10^{19}	$8\times10^{19}(2\times10^{20})$	$2\times10^{20}(1.5\times10^{21})$	1.5×10^{12}
30	4×10^{-7}	0.2	2.4×10^{15}	4×10^{13}	2×10^{15}	$3\times10^{15}(8\times10^{15})$	$7\times10^{16}(5\times10^{17})$	5×10^{7}
1000	0.4	6×10^{3}	1.3×10^{13}	4×10^{7}	6×10^{10}	$8\times10^{10}(2.5\times10^{11})$	$2\times10^{11}(1.5\times10^{12})$	1.5×10^{3}

[a] The time t is the age of the Universe.

[b] The time τ_{fb} is that characteristic of transitions of free electrons to bound atomic states. This is given for hydrogen atoms as $\tau_{fb}(H)$.

[c] The time τ_{bb} is that characteristic of transitions of free electrons between bound atomic states. This is given for hydrogen atoms as $\tau_{bb}(H)$.

sequently the parameter characterizing the Maxwellian distribution, the temperature, is determined by other processes. The energy loss was calculated above in order to provide an order of magnitude estimate for the relaxation time of the electrons.

Conversely, one can find, in terms of the relaxation times, the size of the energy injection necessary for the maintenance of a high electron temperature. It is

$$Q = \epsilon_p \sum \frac{1}{\tau} = 3nkT\left(\frac{1}{\tau_c} + \frac{1}{\tau_{ff}} + \frac{1}{\tau_{fb}} + \frac{1}{\tau_{bb}} + \frac{1}{t}\right).$$

Note that the electron temperature exceeds the radiation temperature by a large factor. In this formula τ_{el} does not enter because the interaction of the electrons with one another redistributes their energy but does not change it. On the other hand, t^{-1} does enter (it would be more accurate to use $2H^{-1}$, where H is the instantaneous value of the Hubble parameter) since the adiabatic cosmological expansion also lowers the plasma temperature.

Thus, the main process in each case is the one for which the respective value of τ is minimal. In the absence of sources of energy, the plasma temperature falls not to zero but to the radiation temperature T_r.

For the cases represented in Table 8, recombination does occur and all of the parameters change. Especially for the highest radiation temperature, however, the question of a stationary electron temperature T_e is of theoretical interest. If the radiation field is a Planck field at the temperature θ, then T_e identically equals θ.

Early in the development of quantum theory, similarities proved to Einstein that radiation just supports nonrelativistic particles in Brownian motion with the corresponding kinetic energy. The quantum picture and the existence of photon momentum were thereby confirmed. Now we consider this self-evident and are interested in a more complicated situation: What will the electron temperature be in an arbitrary (not Planck) radiation field? The answer is easily given if bremsstrahlung processes predominate. Let us denote the spectral volume coefficient of the radiation from the plasma by $Q(\nu, T_e)$. To determine Q, we assume that the plasma is optically thin and that no radiation from outside is incident on it. If now the plasma is located in a radiation field, characterized by the spectral intensity $F(\nu)$ (in the units ergs $[\mathrm{cm}^2\ \mathrm{s\ sr\ Hz}]^{-1}$), then

$$\frac{d\epsilon}{dt} = -\int Q(\nu, T_e)\left[1 - \frac{F(\nu)}{F_{eq}(\nu, T_e)}\right]d\nu\,. \tag{8.3.16}$$

Here $F_{eq}(\nu, T_e)$ is the equilibrium flux corresponding to the electron temperature T_e,

$$F_{eq}(\nu, T_e) = \frac{2h\nu^3}{c^2}\left[\exp\left(\frac{h\nu}{kT_e}\right) - 1\right]^{-1}; \tag{8.3.17}$$

the electron temperature itself is determined from the condition $d\epsilon/dt = 0$.

Clearly, if $F(\nu) = F_{eq}(\nu, \theta)$, then $T_e = \theta$ holds independent of the detailed form of Q. This is so by virtue of the equality

$$1 - F_{eq}(\nu, \theta)/F_{eq}(\nu, T_e) \equiv 0 \quad \text{for} \quad T_e = \theta .$$

If $F \neq F_{eq}$, the answer must be found numerically. It is important that at low frequencies, $F_{eq} \sim \nu^2$ and $Q \sim \ln (kT/h\nu)$; consequently $F(\nu)$ enters into the integral with the weight $\nu^{-2} \ln (kT/h\nu)$, which is large for small frequencies. Hence the intensity indeed determines the electron temperature for low frequencies. The problem is easily generalized to the case of an external source of energy (cosmic rays or plasma oscillations). The new term must then be placed on the right side of the equation for $d\epsilon/dt$.

In the general case of a nonequilibrium spectral intensity $F(\nu)$, the bremsstrahlung processes by themselves do not have to lead to a Maxwellian distribution for the electrons. On the other hand, collisions between electrons will thermalize their distribution. The problem of the interaction of the electrons with the radiation due to scattering (in the limit of low number density, n_e) is more complicated. It was examined by Dreicer (1964). The resulting electron temperature in a nonequilibrium radiation field was given completely by Pero (1968) and by Zel'dovich and Levich (1970). It has the form

$$T_e = \frac{c^2}{2k} \frac{\int F(\nu)[F(\nu) + 2h\nu^3/c^2]d\nu/\nu^2}{\int F(\nu)d\nu} . \tag{8.3.18}$$

The formula looks better if, instead of the flux, one introduces the photon occupation number in phase space. Then

$$kT_e = \frac{h^2 \int n(\nu)[n(\nu) + 1]\nu^4 d\nu}{4h \int n(\nu)\nu^3 d\nu} . \tag{8.3.19}$$

The demoninator is proportional to ϵ, the total energy density of the radiation, because the denominator characterizes Compton energy losses of an electron. The integral in the numerator is the rate of energy gain. The expression for the energy gain will appear once again in the following section. Notice the usual form of the numerator. It contains the factor F^2/ν^2. Consequently, if $F \sim \nu^\alpha$ with $\alpha < 0.5$, then the integral diverges as $\nu \to 0$: a small energy flux at long wavelengths is very important for the electrons even if this flux is negligible with respect to the total energy flux. From the second formula one can see that at long wavelengths (small ν) the induced Compton effect is important. Induced processes occur because the photons are bosons, so their probability of transition into a definite state varies as $n + 1$, where n is the number of photons already in this state.

The classical model of the induced scattering responsible for the heating of the electrons consists first of an electromagnetic wave (wave 1) that forces an electron to oscillate according to the equation

$$m(d\mathbf{x}^2/dt^2) = e\mathbf{E} \quad \text{or} \quad (d\mathbf{x}/dt) = -i(e/m\omega)\mathbf{E}_1$$

when $E_1 \sim \exp(i\omega t)$. Second, another wave (wave 2) acts on the electron by means of the Lorentz force arising from the magnetic field of this second wave:

$$f = (e/c)(dx/dt) \times H_2 = (e^2/mc\omega)E_1 \times H_2 .$$

It is important that H_1, the magnetic field of the first wave, have the same phase as E_1 so that (dx/dt) is shifted by 90° with respect to H; then $\langle (dx/dt) \times H_1 \rangle = 0$. Now if E_1 and H_2 are parallel, there is no such force f. It is thus necessary that the second wave have another direction. In a chaotic field, the squares of separate force impulses are summed so that there occurs a pure energy gain proportional to $|E_1|^2|H_2|^2/\omega^2$. For isotropic radiation, this gives, after the appropriate integration,

$$\int F^2 \frac{1}{\omega^2} d\omega , \qquad F \sim |E|^2 \sim |H|^2 .$$

The term with h disappears in the classical limit. For the Planck spectrum, one again obtains the identity $T_e = T_r$. The important conclusion is that Compton scattering leads to a Maxwellian distribution even in the absence of electron collisions $e + e' \rightarrow e'' + e'''$. It is moreover evident that electron collisions do not impair this. Compton heating of the electrons is important not only in cosmology, but also for the behavior of a plasma in the neighborhood of quasars and pulsars. (For details, see the article by Levich and Sunyaev 1971.) The effect of induced Compton scattering on the spectra of radio sources was examined by Sunyaev (1971).

From a comparison of the characteristic times τ_c and τ_{ff} with the cosmological time t in Table 8, we see that Compton scattering is important for electrons when $z > 10$. After this time, the electrons do not interact with the radiation. Their temperature from that moment until the present time depends on heating by shock waves and cosmic rays and cooling due to adiabatic expansion connected with the general expansion of the Universe. Subsequent processes are examined by Ginzburg and Ozernoy (1965). The characteristic time for the adiabatic expansion is given by t in Table 8. For small z this time is a minimum, which also signifies that any given process can become dominant.

8.4 The Effect of the Electrons on the Radiation Spectrum

From the discussion of the preceding section, we see that Compton scattering is the most important process affecting the electrons for $z \geq 1000$, i.e., for just those times when the plasma is completely ionized. Evidently this process is just as important for photons too.

In the first approximation, scattering from an electron at rest does not change the photon frequency. More precisely, however, ν' (the frequency after scattering) is somewhat less than ν (the frequency before scattering), with the average decrease in frequency satisfying $(|\nu' - \nu|)/\nu = h\nu/mc^2$.

If, on the other hand, the electron is initially moving with velocity v, then the Doppler effect makes the frequency of the scattered photon change within the limits $|(\nu' - \nu)/\nu| \leq v/c$. Thus, scattering from moving electrons leads to diffusion of the photons with respect to frequency. By paying attention to the exact angular distribution and to the presence of induced scattering, one can obtain a differential equation for the spectrum of those photons that are interacting with a Maxwellian distribution of electrons. Using the occupation number density $n(\nu, t)$ to characterize the photons, this equation is

$$\frac{\partial n}{\partial t} = \frac{\sigma_c n_e}{m_e c} \frac{h}{\nu^2} \frac{\partial}{\partial \nu} \left\{ \nu^4 \left[n(n+1) + \frac{kT_e}{n} \frac{\partial n}{\partial \nu} \right] \right\} . \tag{8.4.1}$$

The distribution is assumed to be isotropic.

We will now study several properties of this equation, which was deduced by Kompaneets (1956) and Weymann (1965) and was examined by Zel'dovich and Sunyaev (1969) and by Sunyaev and Zel'dovich (1970*b*, *c*). First, one can rewrite the equation using the dimensionless variables $x' = h\nu/kT_e$ and $dy = \sigma_c n_e (kT_e/m_e c^2)\, cdt$. (Below it is assumed that T_e does not depend on time.) The result is

$$\frac{\partial n}{\partial y} = \frac{1}{x'^2} \frac{\partial}{\partial x'} \left\{ x'^4 \left[n(n+1) + \frac{\partial n}{\partial x'} \right] \right\} . \tag{8.4.2}$$

It is now easy to verify that the Planck form for n, i.e., $n = (e^{x'} - 1)^{-1}$, gives $\partial n/\partial y = 0$. The same is also true in the case of the Bose-Einstein distribution $n = (e^{x'+\mu/kT} - 1)^{-1}$.

Another rather evident property is the conservation of the total number of photons:

$$\frac{dn_\gamma}{dt} = \frac{d}{dt} \int n\nu^2 d\nu = \int \frac{\partial n}{\partial t} \nu^2 d\nu = 0 . \tag{8.4.3}$$

The rate of change of the radiation energy density is given by the formula

$$\begin{aligned} \frac{d\epsilon}{dt} &= \frac{d}{dt} [ah \int n\nu^3 d\nu] \\ &= \frac{4kT_e}{m_e c} \sigma_c n_e ah \int n\nu^3 d\nu - \frac{\sigma_c n_e}{m_e c} ah^2 \int n(n+1)\nu^4 d\nu . \end{aligned} \tag{8.4.4}$$

Now the increase in the radiation energy per unit volume equals the energy given off by the electrons in that volume. If the electrons do not have other sources of energy, their temperature is stationary only if $d\epsilon/dt = 0$. This condition again leads us to formula (8.3.19). The kinetic equation for the photon distribution was written without taking account of the expansion. This is corrected by making the replacement $\partial n/\partial t \to dn/dt + H\nu \partial n/\partial \nu$ (where H is the Hubble parameter) on the left-hand side of the equation. Without Compton scattering, n is conserved even as the frequency changes due to the expansion-generated redshift according to $d\nu/dt = -H\nu$. A simple transformation gives another way to take account of the expansion.

Introduce the radiation temperature which depends on the time according to the redshift law $T_r(t) = T_0(1 + z)$, where T_0 is the temperature today ($T_0 = 2.7$ K). Now let us determine $x = h\nu/kT_r$. The equation for $n(x, t)$ is

$$\frac{\partial n}{\partial t} = \frac{\sigma_c n_e k T_r(t)}{m_e c} \frac{1}{x^2} \frac{\partial}{\partial x} \left\{ x^4 \left[n(n+1) + \frac{T_e}{T_r} \frac{\partial n}{\partial x} \right] \right\} . \qquad (8.4.5)$$

If $T_e = T_r$, the equilibrium form $n = (e^x - 1)^{-1}$ turns out to be an exact solution even though n, as a function of ν and t, depends on time through the time dependence of T_r. The solution is self-consistent because $T_e = T_r$. Finally, this is also the solution for the electron temperature if no energy at all is released in the system.

Now as was promised in § 8.1, we will discuss the question of the release of energy during the various evolutionary stages.

8.5 Early Energy Release and Quasi-Equilibrium

Imagine some small-scale, local perturbations of the homogeneous, isotropic cosmological model during the RD era. Examples of such perturbations can be acoustic waves propagating through the RD plasma or small islands of plasma consisting of antimatter and photons surrounded by the normal plasma of matter and photons. These perturbations eventually disappear. *Why?* The acoustic waves will attenuate due to viscosity, while the antimatter annihilates. But the disappearance of the perturbations is accompanied by the release of energy. Thus, the energy density ϵ_r increases.

Were this to occur in an insulated container, of constant volume, the result would soon be evident. A new state of thermodynamic equilibrium would arise corresponding to the new, increased value ϵ'_r. However, in an evolving, expanding system, equilibrium may fail to be reestablished. For example, at recombination the interaction of the radiation with the electrons ceases, and the photons are essentially free. In fact, just that spectrum of the radiation that exists at recombination is frozen in. In other words, the general cosmological expansion by itself changes the spectrum while preserving its form; i.e., a nonequilibrium spectrum remains nonequilibrium.

Thus, to determine the fate of the perturbation it is necessary to trace the changes in the spectrum with time after the energy release and until recombination. Now, as a consequence of the expansion, the interaction time for the radiation and for the electrons becomes finite. Integrals of the form $\int n_e dt$, $\int n_e T_e dt$, and $\int n_e^2 f(T) dt$ enter the theory. Since $n_e \sim t^{-3/2}$ and $T_e \sim t^{-1/2}$, the values of these integrals depend most strongly on the lower limit of integration, i.e., on the time of energy release (or on the corresponding redshift z), and not on the upper limit—the time of recombination. Hence, we shall examine the sequence of events at the time of and after the energy release. We assume that before the energy release, the spectrum is Planckian.

The final result of annihilation is the production of fast electrons. The scattering of equilibrium radiation by such electrons shifts (more often upward with respect to temperature than downward) the Planck spectrum. One can represent the consequent spectrum as a superposition of Planck spectra with different temperatures:

$$F(\nu) = P(\nu, T)R(T)dT\,, \tag{8.5.1}$$

where $P(\nu, T)$ describes (in units of ergs cm^{-3} Hz^{-1}) a true Planck spectrum having a temperature T (Zel'dovich, Illarionov, and Sunyaev 1972).

Now consider acoustic waves. In the RD plasma, motion with large velocity gradients occurs. In macroscopic terms, the plasma viscosity gives rise to damping of the motion. From the microscopic point of view, an observer astride any electron in the plasma "sees" radiation with diverse redshifts or blueshifts in various directions; i.e., he sees radiation at various "temperatures." The scattering by the electrons then mixes up the radiation incident at different angles.

After the attenuation of the macroscopic motion, the radiation becomes isotropic; its spectrum no longer depends on direction. However, the resultant spectrum is not an equilibrium one, but is again described by a superposition of Planck spectra as in formula (8.5.1) above. Such a spectral form is universal for the first stage after energy release. The time needed for such a spectrum to develop is the characteristic time for Thomson scattering—of order $\tau_T = (\sigma_T n_e c)^{-1}$. Also, inasmuch as each scattering changes the direction of a photon by $\pi/2$ on the average, anisotropy decreases twice as fast.

Suppose that before the energy release there is a Planck distribution at the temperature T_1. From (8.5.1), this corresponds formally to $R(T) = \delta(T - T_1)$, so $\int R(T)dT = 1$. Then, after the energy release, the photon density

$$n_\gamma = A\int T^3 R(T)dT\,, \qquad A = (0.244)\frac{8\pi^3}{c^3}\left(\frac{k}{h}\right), \tag{8.5.2}$$

and also the quantity

$$\int R(T)dT = 1 \tag{8.5.3}$$

are conserved.

Now the energy density may be written as

$$\epsilon_r = A'\int T^4 R(T)dT\,, \qquad A' = \frac{8\pi^5 h}{15c^3}\left(\frac{k}{h}\right)^4. \tag{8.5.4}$$

For low frequencies, the spectral density of the radiation energy is given by the Rayleigh-Jeans formula

$$\epsilon_r(\nu) = \frac{8\pi k}{c^3}\nu^2 T_{RJ}\,, \qquad T_{RJ} = \int T R(T)dT\,. \tag{8.5.5}$$

One can then show that, when n_γ and $\int R(T)dT$ are conserved, a decrease in T_{RJ} accompanies an increase in ϵ_r. This means that the energy release is

accompanied by a transfer of photons from the low-frequency region to the high frequency region. At the same time, if

$$\epsilon'_r = \epsilon_r(1 + \delta) = A'T_1^4(1 + \delta) , \tag{8.5.6a}$$

then

$$T_{RJ} = T_1(1 - \delta/2) . \tag{8.5.6b}$$

The appropriate spectrum for this case is shown in Figure 36 by the dashed line. The solid line represents a Planck spectrum corresponding to the temperature T_{RJ}, and is presented for comparison. To account for the expansion in the above formulae, one must replace T_1 by T_1 $[(1 + z)/(1 + z_1)]$, i.e., by that temperature which, at the given instant (given z), would obtain in the unperturbed case.

The change in the spectrum shown in Figure 36 is very difficult to observe. To do this, exact measurements are necessary in two regions—the Rayleigh-Jeans region ($h\nu < kT$) and the region near the maximum ($h\nu > 2kT$). However, processes occurring subsequently, but still during the RD era, produce much more vivid changes in the spectrum. The point is that the dashed spectrum is not one in thermodynamic equilibrium, and the electrons in the RD plasma are more sensitive to this nonequilibrium character than are our present-day spectral receivers.

The formulae of the preceding section allow one to calculate the temperature which the electrons acquire upon interacting with the radiation whose spectrum is given by the superposition of Planck functions. To a good approximation,

$$T_e \approx \frac{\int T^5 R(T)dT}{\int T^4 R(T)dT} . \tag{8.5.7}$$

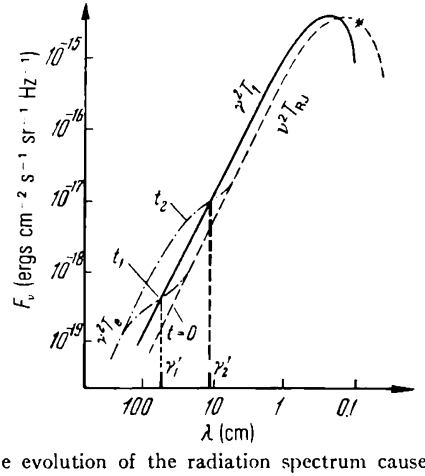


FIG. 36.—The time evolution of the radiation spectrum caused by bremsstrahlung emission in a plasma with a temperature $T_e > T_{RJ}$. The evolution is shown for the time interval $0 < t_1 < t_2$; and the initial spectrum (*dashed curve*) at $t = 0$ is a superposition of Planck spectra. For comparison purposes, the graph exhibits a Planck spectrum characterized by the temperature T_1 that exists before the release of energy (*solid curve*).

The calculation gives (for small δ)

$$T_e \approx T_1\left(1 + \frac{3\delta}{2}\right) \approx T_{RJ}(1 + 2\delta) . \tag{8.5.8}$$

The electrons are hotter than the low-frequency radiation (described by the left-hand dot-dashed line in Fig. 36) and colder than the high-frequency radiation. Moreover, the brightness temperature of the radiation along the dashed curve is not constant; it grows to the right.

Two types of processes occur under these conditions: (1) bremsstrahlung and (2) frequency-changing scattering processes. The integrated effects (over the whole spectrum) of bremsstrahlung emission and absorption are small. In the region of low frequencies, however, its role is important. In fact, the bremsstrahlung spectrum is flat, while the Rayleigh-Jeans spectral function, F_ν, varies as ν^2. Therefore, the time required for the achievement of equilibrium is proportional to ν^2. There is always a region $\nu < \nu'$, with the value of this frequency ν' increasing with time, in which equilibrium corresponding to the temperature T_e is achieved. Figure 36 also shows schematically the evolutionary spectrum (superposition of Planck spectra) resulting from these bremsstrahlung processes. A second Rayleigh-Jeans region arises for $\nu < \nu'$. The step at which the shift occurs moves to the right as time evolves. The small size of the bremsstrahlung emission is responsible for the failure of ν' to reach the maximum of the Planck curve, i.e., $h\nu' \ll kT$ always. In other words, there always remains a substantial spectral region in which the fundamental process is scattering, accompanied by frequency changes (and not bremsstrahlung).

Now let us pose the problem of finding the equilibrium radiation spectrum for a fixed photon density in the limiting case of complete neglect of bremsstrahlung processes. The solution is given by the generalized Bose-Einstein formula with positive chemical potential,

$$n = \left[\exp\left(\frac{h\nu + \mu}{kT_e}\right) - 1\right]^{-1} . \tag{8.5.9}$$

It is easy to convince oneself that $\partial n/\partial t = 0$ results upon substituting this expression for n into the Kompaneets formula, i.e., equilibrium is actually achieved with respect to scattering processes. The spectrum (8.5.9) differs quite sharply from a Planck spectrum in the region $h\nu < \mu$. Asymptotically, we find $F_\nu = \text{const.}\ \nu^3\ (e^{\mu/kT} - 1)^{-1}$ instead of $F_\nu \sim \nu^2$, so the power of ν is changed. Spectrum (8.5.9) is shown as curve *d* in Figure 37 next to a Planck spectrum.

The quantity μ is readily expressed in terms of the energy release. Integrating over the spectrum, we find

$$\begin{aligned} n_\gamma &= \text{const.}\ T^3(1 - 1.37\mu/kT) , \\ \epsilon_\gamma &= \text{const.}\ T^4(1 - 1.11\mu/kT) . \end{aligned} \tag{8.5.10}$$

From this, with $n_\gamma = \text{const} = n_{\gamma,1}$ and $\epsilon_\nu' = \epsilon_{\nu,1}(1 + \delta)$, we obtain

$$T = T_1(1 + 0.64\delta), \quad \mu = 1.4kT\delta. \tag{8.5.11}$$

It was assumed that $\mu/kT \ll 1$ (and, hence, $\delta \ll 1$).

In reality, two mechanisms simultaneously influence the spectrum: bremsstrahlung processes and a redistribution of the photons upon scattering. As a result, spectrum *a* shown in Figure 37 is obtained. It has a peculiar dip at a definite frequency ν_c. At a smaller frequency, equilibrium is maintained by bremsstrahlung processes. In the transition region $\nu \sim \nu_c$, photons diffuse toward increasing frequency. For $\nu > \nu_c$, Bose-Einstein equilibrium holds and is characterized by an appropriate chemical release; a decrease in the chemical potential occurs. This happens as a consequence of the inflow of photons created with frequencies $\nu < \nu_c$ into the region $\nu > \nu_c$. Meanwhile, the quantity ν_c remains virtually constant. The depth of the dip decreases, i.e., the spectral distortion smears out (see curve *b* in Fig. 37), and the spectrum approaches a true Planck equilibrium form (*c* in Fig. 37):

$$\epsilon_\nu' = \epsilon_\nu(1 + \delta) = \sigma T'^4, \quad T' = T_1(1 + \delta/4). \tag{8.5.12}$$

A detailed account of all of these aspects appears in the studies by Sunyaev and Zel'dovich (1970*b*) and by Illarionov and Sunyaev (1974*a*, *b*)

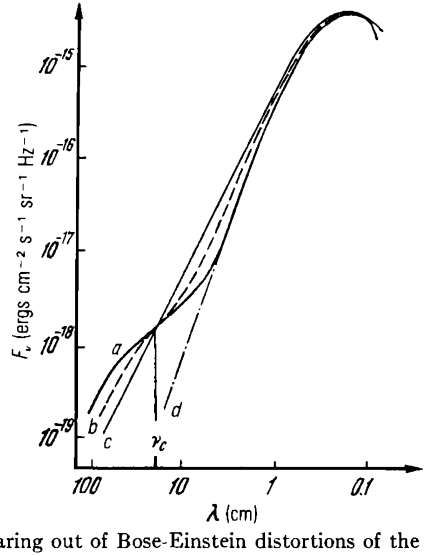


FIG. 37.—The smearing out of Bose-Einstein distortions of the radiation spectrum by the joint action of Comptonization and bremsstrahlung. The evolution takes place via a sequence of spectra *a*, *b*, *c*, which are Bose-Einstein at short wavelengths and have successively decreasing chemical potentials ($\mu_a > \mu_b > \mu_c = 0$). The final spectrum, *c*, is blackbody. One can see the effects of bremsstrahlung by comparing the initial (bremsstrahlung plus Comptonization) spectrum *a* with a pure Bose-Einstein (pure Comptonization) spectrum *d*, which has the same chemical potential as *a*: $\mu_d = \mu_a$.

and in the review by Sunyaev and Zel'dovich (1980). Here we confine ourselves to listing the results:

1. The greatest distortions in the spectrum occur in the region $h\nu < kT$, i.e., in the Rayleigh-Jeans part of the spectrum. This region corresponds to wavelengths greater than a critical wavelength λ_m which depends weakly on the time and amount of energy released. The region is moreover determined completely by the electrons; i.e., λ_m is expressed solely in terms of Ω:

$$\lambda_m \propto 3\Omega_0^{-7/8}, \quad \text{so } \lambda_m = 5.8 \text{ cm}, \quad \Omega_0 = 1; \qquad \lambda_m = 55 \text{ cm}, \quad \Omega_0 = 1/45 \tag{8.5.13}$$

(where Ω_0 refers to the case when $H_0 = 75$ km s^{-1} Mpc^{-1}).

2. The sensitivity of the spectrum to energy release depends substantially on the time of the release. The maximum sensitivity is quite large; in order to obtain a dimensionless deviation α from the Rayleigh-Jeans law for $\lambda = \lambda_m$, it is necessary that the chemical potential be $\mu = 2\alpha\, h\nu_m$, where $\nu_m = c/\lambda_m$. At the same time the dimensionless energy release δ (referred to the energy density ϵ_r at the instant of the release) is

$$\delta = 1.4\alpha \frac{h\nu_m}{kT} = 0.25\Omega_0^{7/8}\alpha; \tag{8.5.14}$$

i.e., 0.25α for $\Omega_0 = 1$ or 0.013α for $\Omega_0 = 1/45$.

3. At the present time, deviations from the Rayleigh-Jeans law are not observed. Conservative estimates of the accuracy of the measurement give $\alpha < 0.1$, which corresponds to $\delta < 0.025$ for $\Omega_0 = 1$ and $\delta < 10^{-3}$ for $\Omega_0 = 1/45$. These bounds are quite sharp and substantially affect the choice of possible types of spectra of perturbations to the homogeneous, isotropic model.

4. The most "dangerous" period, i.e., the one giving the greatest changes α in the spectrum for a given energy release δ, is that with redshift z in the interval

$$2.7 \times 10^4 \Omega_0^{-1/2} < z < 1.2 \times 10^5 \Omega_0^{-6/5}; \tag{8.5.15}$$

that is,

$$2.7 \times 10^4 < z < 1.2 \times 10^5 \quad \text{for} \quad \Omega_0 = 1,$$

$$1.8 \times 10^5 < z < 1.2 \times 10^7 \quad \text{for} \quad \Omega_0 = 1/45.$$

For a much later energy release the evolution time for the spectrum is insufficient to allow for the development of deviations away from the Rayleigh-Jeans spectrum. A late energy release (in particular, one occurring after the RD period) receives special treatment in the next section. A much earlier energy release, occurring at a value of z greater than those indicated above, allows the deviations to be smeared out. The upper limit on z in formula (8.5.15) corresponds to a decrease in the sensitivity by about a factor of e in comparison to the maximum sensitivity.

5. For a large energy release, the smearing occurs slowly. The number of bremsstrahlung photons which must be produced to restore equilibrium

ε proportional to δ, while the emission rate depends weakly on δ. The time required for the establishment of equilibrium is proportional to δ, and not to $\ln \delta$ as is the case in the linear theory. For $\delta \sim 1$, the deviations in the spectrum corresponding to $\alpha \sim 0.1$ can be noticed up to $z \sim 10^7 \Omega_0^{-4}$. This gives $z = 10^7$ for $\Omega_0 = 1$ and $z = 10^{10}$ for $\Omega_0 = 0.2$. But for $z = 10^9$, the number of electron-positron pairs is already many times greater than the present day number of electrons per unit of comoving volume. Therefore the critical value of z realistically achieves a maximum of 10^8 for $\Omega_0 \approx 0.7$ and does not grow with any further decrease in Ω_0.

8.6 When Does the Relic Radiation Arise?

Detailed research over many years on the evolution of the spectrum has led to conclusions important for the theory of the hot Universe. Most significantly, an answer has been obtained to the simple question: When does the equilibrium relic radiation arise? From the numbers presented it follows that the radiation equilibrium is already substantial for z greater than 10^7–10^8 (for the two Ω_0 variants).

In the theory of the homogeneous model, there is a period when other forms of energy are transformed into radiation. The earliest such instant is that of the annihilation of electrons and positrons. Can one hope for observations of the spectral deviations originating at that time?

The energy release is large: in the interval from $T = m_e c^2$ to $T = 0.1 m_e c^2$ (corresponding to $2 \times 10^9 > z > 2 \times 10^8$) annihilation more than doubles the radiation energy. Thereupon, the energy release rapidly falls and yields $\delta = 3 \times 10^{-3}$ in the interval $2 \times 10^8 > z > 10^8$; later, $\delta = 2 \times 10^{-7}$ for the entire remaining interval $5 \times 10^7 > z$. Moreover, larger values of δ cannot be observed for any value of Ω_0, because the excess electrons and positrons participate in the smearing of the spectrum.

The energy release for nuclear reactions (the formation of ^{4}He) takes place at $t \sim 100$ s, or redshift $z \sim 4 \times 10^8$, when the temperature is $T \sim 100$ keV. The energy is of the order of 28 MeV for the ^{4}He nucleus. When scaled for each nucleon and keeping in mind that only 30% of the matter becomes ^{4}He, this gives about 3 MeV per nucleon. This energy corresponds to the quite small value, $\delta \sim 3 \times 10^{-7}$. Moreover, by taking account of the still large number of electrons and positrons present at this time, one sees that the release of nuclear energy cannot produce noticeable distortions of the spectrum observed today.

8.7 Late Energy Release

The late energy release case was examined by Zel'dovich and Sunyaev (1969). If energy is released at a late stage after recombination, then re-ionization occurs. The electron temperature increases to a value greater than 10^4 K. Nevertheless, the radiation temperature falls from 4000 K at

the beginning of the recombination to the present-day value of 2.7 K. The late stage is thus characterized by the condition $T_e \gg T_r$ (instead of $T_e - T_r \ll T_r$ as in § 8.5). In the basic equation (8.4.5), the term multiplied by T_e/T_r remains, while the term $n(n+1)$ is thrown away because of its small size. In this way we obtain the linear equation

$$\frac{\partial n}{\partial t} = \frac{\sigma_c n_e c k T_e}{m_e c^2} \frac{1}{x^2} \frac{\partial}{\partial x}\left[x^4 \frac{\partial n}{\partial x}\right]. \tag{8.7.1}$$

Here x refers to the radiation temperature, $T_r(1+z)$, where T_r is defined as the radiation temperature that would be obtained today if there were no energy release. Thus we allow $T_r \neq 2.7$ K.

We introduce the dimensionless quantity y satisfying

$$dy = (\sigma_c n_e k T_e/m_e c)\, dt = (k T_e/m_e c^2)\, d\tau$$

with $d\tau = \sigma_c n_e c dt$. Here τ is the optical depth. After integration, y becomes

$$y = \int_{t_1}^{t} \frac{k T_e(t)}{m_e c^2}\, d\tau\,, \qquad \tau = \tau(t)\,, \tag{8.7.2}$$

where, at t_1, the electron temperature is increased due to the energy release. The equation is not applicable when $T_e \approx T_r$.

The present spectrum depends only on the quantity obtained from an integration to $t = t_0$. Various functions $T_e(t)$ give the same final spectrum if the same value y_0 is obtained from (8.7.2). This dependence on one parameter reduces uncertainties and facilitates comparisons with observations.

Proceeding, we are first and foremost interested in the present spectrum corresponding to y_0 (in the subsequent development, we will omit the subscript zero on y_0). The solution of equation (8.7.1) takes the form

$$n(x, y) = (4\pi y)^{-1/2} \int_0^\infty n(\xi, 0) \exp\left[\left(-\ln\frac{x}{\xi} + 3y\right)^2/4y\right] \frac{d\xi}{\xi}. \tag{8.7.3}$$

The basic properties of the solution are

$$\begin{aligned} n_\gamma &= \int n(x, y) x^2 dx = \int n(x, 0) x^2 dx = \text{const.}\,; \\ \epsilon_\gamma &= \int n(x, y) x^3 dx = e^{4y} \int n(x, 0) x^3 dx = \epsilon_0 e^{4y}\,. \end{aligned} \tag{8.7.4}$$

Solution (8.7.3) is cumbersome. In the subsequent development, we shall thus assume as an initial condition the existence of an equilibrium spectrum $n(x, 0) = (e^x - 1)^{-1}$, neglecting, thereby, the case when the energy is released earlier. For $x \ll 1$, $n(x, 0) = x^{-1}$. Setting $n = f(y)/x$, we obtain $f = e^{-2y}$, so that $n(x, y) = e^{-2y}/x$ for $x \ll 1$. The condition $x \ll 1$ is valid for the long-wavelength part of the spectrum. If $n(x, y) < n(x, 0)$, then this means that long-wave observers will measure a quantity lower than the unperturbed temperature:

$$T_{RJ} = T_r e^{-2y}\,. \tag{8.7.5}$$

We already know that $T_{RJ} = 2.7$ K. Consequently, in order to compare the theory with the observations, we must take the corresponding temperature T, $T_{RJ}e^{12y} = 2.7\,e^{2y}$ K for the particular value of y taken. In Figure 38 are shown curves for several values of y; all correspond to $T_{RJ} = 2.7$ K. For $y = 0$, the curve represents a Planck spectrum. The other curves coincide at the left end, but are higher than the Planck curve at short wavelengths. The total energy density is given by the expression

$$\epsilon_r = \sigma T_r^4 e^{4y} = \sigma T_{RJ}^4 e^{12y} = 4 \times 10^{-13} e^{12y} \text{ ergs cm}^{-3}. \quad (8.7.6)$$

The energy gain equals

$$\Delta\epsilon_r = \sigma T_{RJ}^4 (e^{12y} - e^{8y}) \text{ ergs cm}^{-3}. \quad (8.7.7)$$

Per gram of matter, this gives

$$\frac{\Delta\epsilon}{\rho_m} = \frac{4 \times 10^{-13}(e^{12y} - e^{8y})}{2 \times 10^{-29}\Omega_0 h_0^2} = 2 \times 10^{16}(\Omega_0 h_0^2)^{-1}(e^{12y} - e^{8y}) \text{ ergs g}^{-1}. \quad (8.7.8)$$

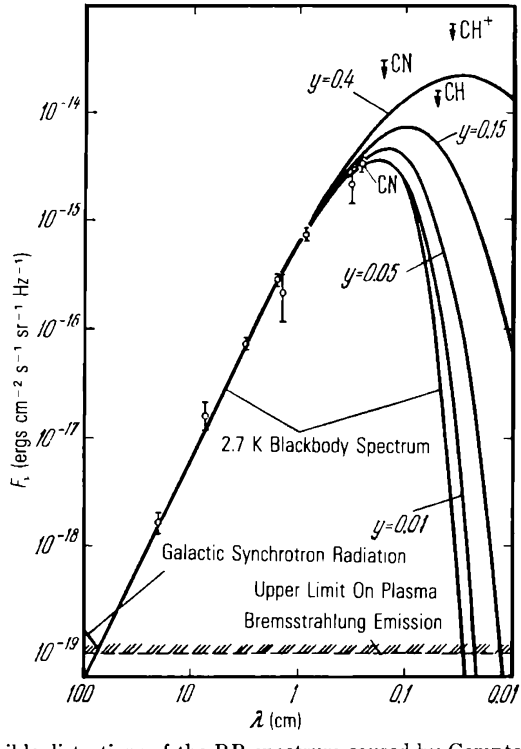


FIG. 38.—Possible distortions of the RR spectrum caused by Compton scattering from hot electrons. Some experimental points are plotted over the RR curve.

But, since the radiation energy per unit mass decreases with the expansion as $1 + z$, the energy released per gram of matter equals

$$\frac{\delta\epsilon}{\rho_m} = 2 \times 10^{16}(\Omega_0 h_0^2)^{-1}(1 + z_*)(e^{12y} - e^{8y}) \text{ ergs g}^{-1}, \qquad (8.7.9)$$

where z_* is the average redshift at which the energy release and the exchange of energy between the electrons and the photons occur.

The optical depth τ is given by the expression (α is the ratio of the density of ionized gas to the total density)

$$d\tau = n_e \sigma_e c dt \approx 0.072(\Omega_0 h_0)\alpha \frac{(1 + z)dz}{(1 + \Omega_0 z)^{1/2}} \approx 0.072(\Omega_0 h_0^2)^{1/2} z^{1/2} dz . \quad (8.7.10)$$

The last equality is valid for $\alpha \approx 1$ and $z \gtrsim \Omega_0^{-1}$. Before proceeding, we shall describe a refined method for calculating the spectral distortions of the microwave radiation.

With the initial spectrum $n_0(x)$, the solution to equation (8.7.1) may be

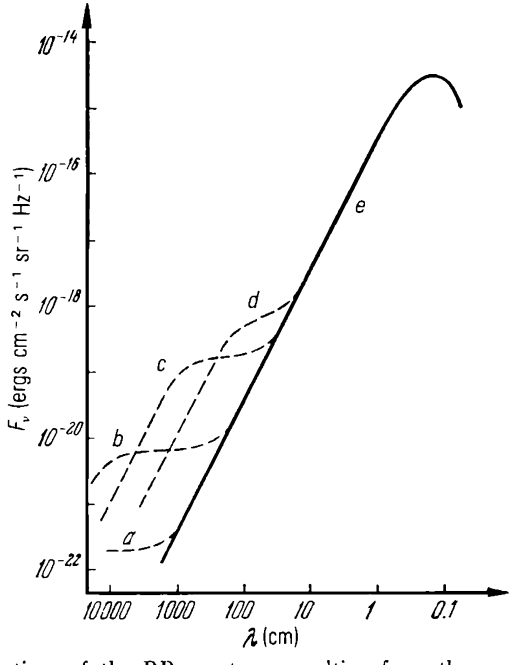


FIG. 39.—Distortions of the RR spectrum resulting from the combined effects of Compton scattering ($y \approx 0.1$) by the electrons of a hot plasma, and the additional bremsstrahlung radiation produced by the hot plasma. The plasma temperature T_e is maintained at a constant level for some interval of the redshift. For the curve labeled *a*, $T_e = 10^8$ K. in the interval $1394 < z < 1400$; for curve *b*, $T_e = 10^7$ K at $1340 < z < 1400$; for curve *c*, $T_e = 10^6$ K at $700 < z < 1400$; and for curve *d*, $T_e = 10^5$ K at $700 < z < 1400$. In the last case, the Compton distortions are small, corresponding to a parameter value $y \approx 0.01$. Curve *e* is the undistorted Planck spectrum.

written as the superposition of shifted spectra:

$$n(x, t) = \int n_0(x/\phi)R(\phi, t)d\phi \,. \tag{8.7.11}$$

In particular, if the initial spectrum is Planckian, a superposition of Planck spectra is obtained. This result is natural in view of the discussion of 8.5. In this case,

$$n(x, y) = \int n_p(\nu/T)R(T, y)dT \,, \tag{8.7.12}$$

where $n_p(\nu/T) = [\exp(h\nu/kT) - 1]^{-1}$. The equation for $R(T, y)$, equivalent to the Kompaneets equation, takes the form

$$\frac{\partial R}{\partial y} = \frac{1}{T^3}\frac{\partial}{\partial T}\left[T^4\frac{\partial}{\partial T}(TR)\right]. \tag{8.7.13}$$

The advantage of shifting attention to R is that the initial conditions are simplified: $R(T, y = 0) = \delta(T - T_0)$; R goes zero for $T \to \infty$ and for $T \to 0$. Two conservation laws then follow:

$$\int T^3R(T, y)dT = \text{const.}\,, \tag{8.7.14a}$$

$$\int R(T, y)dT = \text{const.} \tag{8.7.14b}$$

The first integral is proportional to the photon number density; it reflects photon conservation. The second integral guarantees that the radiation is not diluted.

For an initial condition in the form of a delta function, the solution to (8.7.13) is

$$R(T, y) = (4\pi y)^{-1/2}\exp\left\{4y - \frac{1}{4y}\left[\ln\left(\frac{T}{T_0}\right) + 5y\right]^2\right\}. \tag{8.7.15}$$

The previous conclusions may now be recovered quite easily:

$$T_{\mathrm{RJ}} = \int TR(T, y)dT = T_0e^{-2y}\,, \tag{8.7.16a}$$

$$\epsilon_\gamma \propto \int T^4R(T, y)dT = \epsilon_0e^{4y} \propto T_{\mathrm{RJ}}{}^4e^{12y}\,. \tag{8.7.16b}$$

Now we shall return to the physical consequences of a (relatively) recent heating of the intergalactic medium (IGM). An important consequence of late energy release is connected with the bremsstrahlung emission by the gas. This radiation, however, depends on the time variation of the electron temperature and is not determined by y alone. For high T_e at present, the emission contributes to the diffuse X-ray background. Observations suggest (see Field and Perrenod 1977) that $T_{e0} \lesssim 4.4 \times 10^8$ K. In contrast, suppose that after recombination the electron temperature does not greatly exceed the photon temperature (e.g., $T_e \approx 10^4$ K for $300 \lesssim z \lesssim 1400$). The bremsstrahlung photons produced then now find themselves shifted into the long wavelength part of the spectrum; they may distort the Rayleigh-Jeans part of the spectrum.

From a discussion of the distortions of the microwave spectrum (see

Field and Perrenod 1977; Jones and Steigman 1978), it has been determined that $y \lesssim 0.03$. Assuming prompt heating (at z_*) and adiabatic cooling of the IGM (and assuming the observed, diffuse X-ray background is from a hot IGM), $T_r(z) \approx 4.4 \times 10^8\ (1+z)^2$ K. The corresponding value of y predicted by (8.7.2) is

$$\begin{aligned} y &\approx 5.3 \times 10^{-3}(\Omega_0 h_0) \int_0^{z_*} \frac{\alpha(z)(1+z)^3}{(1+\Omega_0 z)^{1/2}}\, dz \\ &\approx 1.5 \times 10^{-3}(\Omega_0 h_0^2)^{1/2} z_*^{7/2}\,. \end{aligned} \tag{8.7.17}$$

The last equality is valid for $\alpha \approx 1$ and $z_* \gtrsim \Omega_0^{-1}$. For $y \lesssim 0.03$ we obtain, from (8.7.8) and (8.7.17),

$$(\Omega_0 h_0^2)^{1/2} z_*^{7/2} < 20\,, \tag{8.7.18a}$$

$$\left(\frac{\delta\epsilon}{\rho_m}\right)_* < \frac{3.2 \times 10^{15}}{(\Omega_0 h_0^2)}\,(1+z_*) \text{ ergs g}^{-1}\,. \tag{8.7.18b}$$

To match the observed intensity of the X-ray background, Field and Perrenod (1977) require $\Omega_0 h_0^2 \gtrsim 0.12$ (i.e., $\Omega_0 \gtrsim \frac{1}{2}$, $h_0 \approx \frac{1}{2}$). This leads to $z_* \lesssim 3.2$ and $(\delta\epsilon/\rho_m)_* \lesssim 8 \times 10^{16}$ ergs g^{-1}.

For the future, it would be extremely valuable to obtain simultaneous measurements at two wavelengths λ_1 and λ_2 such that $h\nu_1 < kT_r$ and $h\nu_2 \gtrsim 3kT_r$. Such measurements will yield important information concerning the evolution subsequent to recombination ($z \approx 1400$). If such data could be obtained with angular resolution on the order of 0.01 radians or better, there will be the possibility of distinguishing spatial inhomogenieties from spectral distortions.

For a spatial inhomogeneity, the fluctuations in all parts of the spectrum are correlated. In contrast, the distortions due to clouds of hot electrons lead to an increase at high frequencies ($h\nu \gtrsim kT_r$) and a decrease at low frequencies ($h\nu \lesssim kT_r$). The data at present indirectly indicate that the energy transferred to the radiation is small. This means that since the fluctuations are small, accurate (and therefore difficult) experiments will be required.

8.8 The Intergalactic Gas: Its Radiation and Density

The absence of neutral gas in the intergalactic medium and the difficulty in studying the radiation from a hot, ionized, intergalactic gas were indicated in the Introduction. With hot gas at the critical density, the total radiation per electron is shown as a function of temperature in Figure 40. For historical reasons, the calculations here are carried out for $H_0 = 100$ km s^{-1} Mpc^{-1} and $\rho_c = 10^{-29}$ g cm^{-3}. At the end of the section, we will discuss how the conclusions change if $H_0 = 50$ km s^{-1} Mpc^{-1}. One curve corresponds to the case of pure hydrogen, the other to the case of 70% hydrogen and 30% helium. The characteristic maximum at $T = 2 \times 10^4$ K is

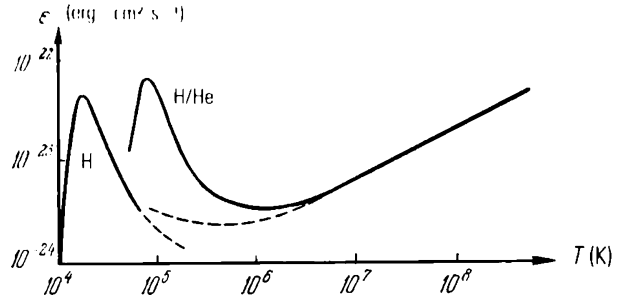


Fig. 10.—The temperature dependence of the radiated energy from an intergalactic hydrogen helium plasma. The abscissa gives the logarithm of the temperature in kelvins, while the ordinate gives the energy emitted by the plasma. The curve marked *H* is for a pure hydrogen plasma; that marked H/He is for a mixture 30% helium by mass.

mainly due to line emission by hydrogen; the maximum at $T \approx 8 \times 10^4$ K arises mainly from line emission of He II (at $\lambda = 304$ Å). For a temperature much higher than this maximum, a large part of the energy is radiated by the bremsstrahlung mechanism. It has a spectrum which is described roughly by the formula

$$F_\nu \sim e^{-h\nu/kT} g(\nu) \,. \tag{8.8.1}$$

Detailed calculations were performed by Gould and Ramsey (1966), by Weymann (1967), and by Doroshkevich and Sunyaev (1969).

Let us make use of the observations of distant quasars in order to estimate the possible amount of neutral hydrogen in intergalactic gas following Gunn and Peterson (1965). If neutral hydrogen were present in noticeable amounts, it would cause Lyman-α absorption in the spectra of quasars. Since absorbing atoms are located over the whole path from a given quasar to us, they possess various Hubble recession velocities. The Doppler effect causes the absorption line to be stretched out into a band. It is sufficient to have a density of neutral H atoms $n_{\rm H} \approx 10^{-10}$ cm^{-3} in order for quasars with redshifts $z \approx 2$ to possess an observable absorption band in the optical part of the spectrum. The absence of such bands in the spectra of quasars gives an upper limit on $n_{\rm H}$. If one believes that the intergalactic gas is plentiful, then he must assume that it has a high temperature and is therefore ionized. In particular, in order to explain the absence of neutral hydrogen for $z \sim 2$ quasars, a temperature $T > 10^6$ K is necessary (Gunn and Peterson 1965). A rough estimate can be made according to the Elvert formula $n_p/n_{\rm H} \approx 10^6 (kT/I_{\rm H}) \exp(-I_{\rm H}/kT)$ which takes into account both collisional ionization and recombination accompanied by radiation. The required temperature can be lower if some of the ionization is due to ultraviolet radiation emitted earlier. Actually, the calculations were done based upon a consideration of the kinematics of ionization and of recombination and using more precise values for the appropriate cross sections. At

low densities, ionization of the gas by radiation can play a substantial role. Under these circumstances, the conclusion concerning the high temperature becomes less certain.

Let us now examine the energy balance. For $z < 2$ and $\Omega_0 < \frac{1}{3}$, cooling occurs mainly as a result of adiabatic expansion. Consequently, without a new release of energy, $T_g \sim (1 + z)^2$ and $T_g > 10^6 [(1 + z)/3]^2$ K. The contribution of various processes to the cooling of the gas is presented in Figure 41 for $\Omega_0 = 1$ and $\Omega_0 = 0.1$.

For the temperature estimated above, the predicted flux falling on the Galaxy generated by gas with density $\Omega_0 = 1$ is shown in Figure 42. The calculations were done by Doroshkevich and Sunyaev in 1969 (see Longair and Sunyaev 1971). Measurements in the optical band (Roach and Smith 1968) give an upper limit for the intergalactic background of $F_\nu < 10^{-19}$ ergs $(\text{cm}^2\,\text{s sr Hz})^{-1}$. This includes the background from our Galaxy and from the solar system (the zodiacal light). But even if one removes this background, the background from the distant galaxies $(0.5 < z < 2)$, indistinguishable even in the best telescopes, still gives much more radiation than does the intergalactic gas at this temperature.

Consequently, the measurements must be conducted not in the optical but in the ultraviolet region. There, the stellar spectra are far weaker. For this it is necessary to take a receiver beyond the limits of the atmosphere; the experiment of Kurt is described in Kurt and Sunyaev (1967*a*, *b*). His measurements were done in the interval 1225–1350 Å; in this region the background is much weaker. The resulting upper limit thus corresponds

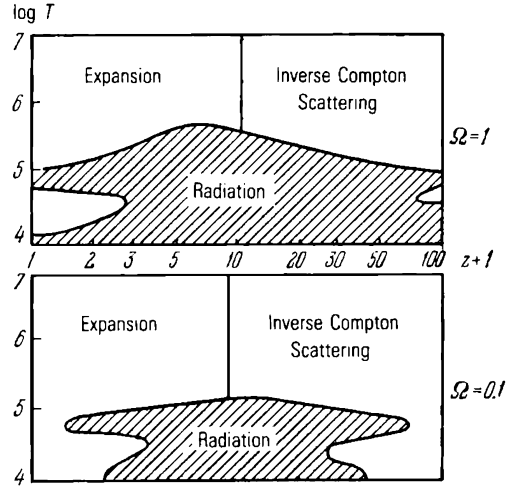


FIG. 41.—The principal cooling process for a hot plasma in the expanding Universe, as a function of the plasma temperature T and redshift z. The upper plots are for $\Omega = 1$; the lower, for $\Omega = 0.1$. In the crosshatched region plasma emission of radiation dominates the cooling. At large redshift Compton cooling dominates.

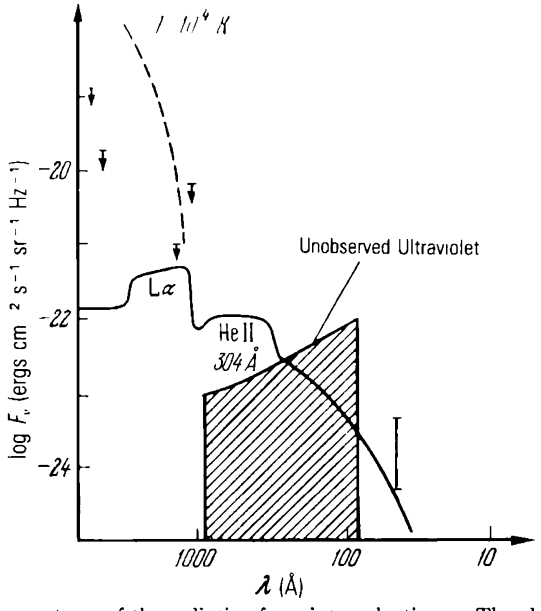


FIG. 42.—The spectrum of the radiation from intergalactic gas. The observational data are shown by arrows (upper limits on the flux) and by a vertical error bar.

to a region of the spectrum where the neutral hydrogen of our Galaxy is transparent. The experiment was performed using a satellite-mounted ionization counter without an anticoincidence circuit. Cosmic rays, therefore, contributed to the measured background. The result for the ultraviolet flux is $F_\nu < 10^{-21}$ ergs (cm^2 s sr Hz)$^{-1}$ (Kurt and Sunyaev 1967*a*, *b*, 1970) which corresponds to $\Omega_0 < 4$. The result is not very significant, but it is possible to improve the method. At the present time though, a different, indirect approach (Sunyaev 1969*b*) gives more definite results.

Observations of the 21 cm radiation from atomic hydrogen can yield the distribution of neutral hydrogen both within and around galaxies, including ours. It is known, for example, that the H I halo often stretches far from the center of a galaxy and consists of very low density H I ($n_H \sim 3 \times 10^{-3}$ to 10^{-2} cm^{-3}). In several cases there are also thin, rarefied gas bridges between neighboring galaxies. Sunyaev noted that such neutral hydrogen is very sensitive to ionization by ultraviolet radiation ($h\nu > I_H$). Order-of-magnitude estimates lead to a lower bound for the lifetime of such halos of order 10^7 years. Thus, the radiation flux must be sufficiently weak. These considerations then give the limit $F_\nu < 10^{-23}$ ergs (cm^2 s sr Hz)$^{-1}$ for $\nu > I_H/h = 3 \times 10^{15}$ s^{-1}.

Together with the calculations of the radiation from the intergalactic gas, this gives a strict limit on the density of the intergalactic gas: $\Omega_0 < 0.3$–0.2 (0.3 for pure hydrogen, 0.2 for 70% hydrogen and 30% helium).

This result once served as a strong argument in favor of the open model

of the Universe. Other components of the matter and energy (the galaxies, the radiation, and the neutrinos and other weakly interacting particles) give a contribution to Ω_0 much less than unity. Indeed, the hot, completely ionized intergalactic gas (more precisely, the gas between clusters of galaxies) is least of all amenable to observations. We thus appreciate the significance of the limit on its density of $\Omega_0 < 0.2$.

Unfortunately, the marked decrease in the Hubble constant makes this result less definite. For a given Ω_0, the density is proportional to H_0^2, the bremsstrahlung emission per unit volume is proportional to $\sim\rho^2 \sim H_0^4$, and the effective length over which the radiation is incident is of the order of c/H_0. Thus, for a specified temperature, the radiation is proportional to $\sim\Omega_0^2 H_0^3$. Hence, for a given, measured amount of radiation and an assumed temperature, $\Omega_0 \sim H_0^{-3/2}$. A decrease in H_0 by a factor of 2—from 100 km s^{-1} Mpc^{-1} to 50 km s^{-1} Mpc^{-1}—leads to an increase in Ω_0 by about a factor of 3. The value of Ω_0 thereby increases to 0.9 for the case of pure hydrogen and to 0.6 for the case of 70% hydrogen and 30% helium. Remember that we are concerned with upper limits; thus, more precise observations and a better theory can lead to a reduction of these quantities. Indeed, Field and Perrenod (1977) note that for $T_{e0} \approx 4.4 \times 10^8$ K, $h_0 = \frac{1}{2}$ and $\Omega_0 \approx \frac{1}{2}$, the observed X-ray background can be accounted for. Concerning Ω_0, we refer the reader to the indirect considerations related to the creation of galaxies and to the growth of perturbations of the Universe (see § 9.6).

Let us return to the radiation from the gas. Suppose that the gas is located inside a cluster of galaxies. This gas apparently is gravitationally bound and only flows out into the space between clusters slowly and in insignificant quantities. Besides, accretion compensates for this outflow of gas. What could be the amount of this inhomogeneously distributed gas?

We shall call the density of the gas in clusters of galaxies, referred to the total volume of space per cluster, Ω_{cl}. Measurements at 21 cm show that neutral hydrogen is absent from the clusters. Quantitatively, $n_{H\,I} < 10^{-7}$ cm^{-3} in the clusters, which corresponds to the average density $n_{H\,I} \lesssim 10^{-9}$ cm^{-3}, i.e., $\Omega_{cl,H\,I} < 10^{-4}$. Now let us turn to the question of the amount of ionized gas in clusters.

We shall denote by α the volume inside a cluster divided by the total volume per cluster. Apparently, α is roughly of the order of 10^{-2}. The actual density of gas in clusters is then $\Omega_{cl}\alpha^{-1}$. The radiation is proportional to the square of the density of ionized gas and to the volume the gas occupies, i.e., to $(\Omega_{cl}\alpha^{-1})^2\,\alpha$. Consequently, if the gas is concentrated in clusters, its contribution to Ω_0 must be even smaller than if it is homogeneously distributed.

Returning to the IGM, we must consider the case when the temperature is much higher than 10^7 K. The radiation in the visible and ultraviolet region decreases ($\sim T_e^{-1/2}$), but the X-ray emission increases. As we have

noted, Field and Perrenod have shown that an IGM at a temperature $T_0 \sim 1.1 \times 10^8$ K and $\Omega_0 \lesssim \frac{1}{2}$ (for $h_0 \sim \frac{1}{2}$) is consistent with the observed X-ray background.

Observations from the *Uhuru* satellite have led to the detection of powerful X-ray sources in some rich clusters of galaxies. The spectrum is apparently consistent with the creation of the X-rays in an optically thin, hot gas. So, for example, the X-ray observations of the Coma cluster of galaxies are explained best of all by the presence in this cluster of a hot, intergalactic gas with a temperature of $T \approx 10^8$ K and a density of the order of 10^{-3} cm^{-3}. Meanwhile, the total mass of the gas in the cluster exceeds the sum of the masses of the galaxies comprising the cluster. It is, however, several times less than the virial mass; i.e., it is insufficient to stabilize the cluster gravitationally. Now the velocity dispersion of the galaxies in this cluster is $\sim$1500 km s^{-1}. Notice that this value corresponds to $v^2/2 \sim kT/m_p$, where $T \approx 10^8$ K is the observed temperature of the gas in the cluster. Consequently, one can imagine that the motion of the gas, at the velocity characteristic of that of the galaxies in the total gravitational field of the cluster, just gives the observed temperature if gas clouds collide.

Notice, too, that in the radio region of the spectrum, the radiation from the hot gas is very weak. Much more significant is the Compton distortion of the microwave spectrum due to its interaction with the hot gas in a cluster. The average photon energy then increases while the low-energy part of the spectrum is depleted so that the Rayleigh-Jeans radiation temperature falls. Correspondingly, observations in the centimeter wavelength band, in the direction of a cluster of galaxies with hot gas, yield a brightness temperature lower than in other directions (Sunyaev and Zel'dovich 1970*a*, 1972*b*). The deviation should be

$$\frac{\Delta T}{T} = -2\sigma_c n_e l \frac{kT_e}{m_e c^2}, \tag{8.8.2}$$

where l is the size of the cluster. Parijsky (1972) has apparently observed this effect.

Recently Gull and Northover (1975) also observed the reduction of the RJ temperature in the direction of X-ray emitting clusters. On the other hand, Lake and Partridge (1977) (see Gull 1976) did not find any effect exceeding the errors. There is another theory for the origin of the cluster X-rays. A much smaller number of ultrarelativistic electrons ($\gamma = E/mc^2 \gg 1$) could give X-rays by scattering of microwave photons; in this case the reduction of the Rayleigh-Jeans part is totally negligible. The measurements of the reduction are crucial to permit a choice between hot gas ($T \sim 10^8 \sim$ 10 keV) or ultrarelativistic particles and also to measure the amount of the hot gas. Observations by Leningrad astrophysicists (Golenetsky and Mazets 1971; Golenetsky *et al.* 1971) suggest that the data obtained by the

Vette group (Vette *et al.* 1970) concerning the γ ray background in this region is sharply overestimated. This is an argument against the ultrarelativistic theory.

8.9 Radio Emission by the Ionized Intergalactic Gas and the Epoch of Neutral Hydrogen

In § 8.2 we studied the process of recombination. Later, in § 8.5 we examined the picture presented by the intergalactic gas in which energy is released. The question arises of what happens if this release of energy interferes with the recombination. Of course, our knowledge of the early processes which can release energy is insufficient to answer it definitively. Nevertheless, it is clear that the question of the energy balance of the gas, of its recombination and ionization, is closely connected with the theory of the growth of present-day structure.

In the theory of adiabatic perturbations, which grow as a consequence of their gravitational instability, $\delta\rho/\rho$ is very small—10^{-3} to 10^{-5}—at recombination. Such density perturbations, plus the velocities corresponding to them, lead to the formation of shock waves and, ultimately, to ionization when $z < 10$. At the time of recombination, however, these small perturbations in no way destroy the evolutionary picture of the plasma developed above.

In the vortex theory, the initial velocities are significant—of the order of $0.01c$–$0.1c$. The collisions of supersonic flows and the ionization of the gas take place no later than the time when $z \sim 100$.

Finally, in the charge-symmetric theories, the release of energy upon annihilation can completely prevent recombination. It is quite important to investigate the consequences of this and of any other assumptions on the actual course of recombination. A comparison of the consequences with observations is essential in deciding among the various theories of the formation of structure.

In this connection, the radio emission by an ionized gas is inestimably important (Sunyaev 1968). It was mentioned that bremsstrahlung photons have a flat spectrum at low frequencies $\nu < kT/h$, while their total power varies as $\sim T^{1/2}$.

Radio emission at low frequencies is proportional to $T^{-1/2}$. Consequently, if someone wants to study all possible regions of gas temperature, then the lower limit on the temperature will be given by a comparison of the calculated radio emission with observations. The upper limit on the temperature can be found from the X-ray emission of the gas or from the calculated energy which is required to compensate for the thermal losses; this was done in previous sections of this chapter.

The most suitable wavelengths for study are found from the condition of minimum intensity of the sum of RR and the radiation from discrete

sources. By calculation, this yields $\lambda \sim 50$ cm. The flat spectrum is not found now, and therefore one can give the upper limit

$$F_{\rm ff} < F_\nu^{\rm obs} < 10^{-19} \text{ ergs (cm}^2 \text{ s sr Hz)}^{-1} \tag{8.9.1}$$

for bremsstrahlung emission, denoted by the subscript "ff" (for free-free emission). By more careful measurements and by selecting directions in the sky in which there are no discrete sources, one can significantly increase the wavelength investigated and lower the corresponding limit $F_\nu^{\rm obs}$. Direct calculations give (we are applying the formula for the integrated radiation field produced by homogeneously distributed, evolving sources)

$$F_{\rm ff} = 5 \times 10^{-20}(\Omega_0^2 h_0^3) \int_0^{z_i} \frac{(1+z)}{(1+\Omega_0 z)^{1/2}} T_e^{-1/2}(z)dz \text{ ergs (cm}^2 \text{ s sr Hz)}^{-1} \tag{8.9.2}$$

$$\approx 5 \times 10^{-20}(\Omega_0 h_0^2)^{3/2} \int_0^{z_i} \left[\frac{z}{T_e(z)}\right]^{1/2} dz \text{ ergs (cm}^2 \text{ s sr Hz)}^{-1} .$$

In obtaining the above, the emissivity (per unit comoving volume) of a hot gas was taken to be

$$\epsilon_{\rm ff}(\nu) = 5.44 \times 10^{-39} \frac{g(\nu)}{T_e^{1/2}} \exp\left(\frac{-h\nu}{kT_e}\right) n_e^2 \text{ ergs (cm}^3 \text{ s sr Hz)}^{-1} . \tag{8.9.3}$$

In the radio region ($h\nu \ll kT_e$) the spectrum is flat; $F_{\rm ff}$ is independent of ν for small ν.

Requiring that $F_{\rm ff}(z_i) < F^{\rm obs}$, we obtain

$$(\Omega_0 h_0^2)^{3/2} \int_0^{z_i} z^{1/2} T_e^{-1/2}(z)dz < 2 , \tag{8.9.4}$$

where z_i is the redshift of reionization. By itself, this condition can be fulfilled for $z_i = z_{\rm rec} \approx 1400$. That is, by sufficiently increasing $T_e(z)$, z_i can be made increasingly larger; for $z_i \to z_{\rm rec}$ there is no recombination. For example, for $T_e(z) = T_i = \text{constant}$, (8.9.4) is satisfied by

$$T_i \gtrsim [\tfrac{1}{9}\Omega_0 h_0^2 z_i]^3 . \tag{8.9.5}$$

For $z_i = z_{\rm rec} \approx 1400$, $T_i \gtrsim 3 \times 10^8 (\Omega_0 h_0^2)^3$.

There are, however, other conditions which are incompatible with such a high temperature. One such condition is that energy losses of the gas by Compton scatterings with the RR cannot exceed W, the total nuclear energy per gram. This yields

$$\int_0^{z_i} \frac{(1+z)^2}{(1+\Omega_0 z)^{1/2}} T_e(z)dz \text{ (ergs g}^{-1}) < 10^{-6} W . \tag{8.9.6}$$

There is another condition which limits the temperature of the gas: Distortions in the RR spectrum, as characterized by the parameter y, must not exceed the experimentally determined limit ($y \lesssim 0.03$). This condition

takes the form

$$(\Omega_0 h_0) \int_0^{z_i} \frac{(1+z)}{(1+\Omega_0 z)^{1/2}} T_e(z)dz < 9 \times 10^{10} y . \qquad (8.9.7)$$

In the above conditions it was assumed that $T_e \gg T_r$ and, thus, these formulae are valid only for $z_i \lesssim 1500$ and $T_e \gtrsim 4000$ K. Indeed, gas temperatures $\gtrsim (1\text{–}1.5) \times 10^4$ K are required (for $T_r \lesssim 4000$ K) to ensure essentially complete ionization.

Let us now pose a mathematical question: What is the best method of using two inequalities of the form

$$\int_0^{z_i} T_e^{-1/2}(z) f(z)dz < a , \qquad \int_0^{z_i} T_e(z) g(z)dz < b , \qquad (8.9.8)$$

for determining the maximum z_i? The functions f and g are known from theory, a and b are given by observations, and z_i gives the time of reionization of the gas (the gas is neutral for $1400 > z > z_i$ and ionized for $z_i > z > 0$).

The problem depends on the choice of the unspecified function $T_e(z)$. It is proposed that one takes a linear combination of the two equations with positive coefficients α and β which we will determine later. Thus

$$\int_0^{z_i} (\alpha T^{-1/2} f + \beta T g)dz < \alpha a + \beta b . \qquad (8.9.9)$$

For any z we choose $T(z)$ by minimizing the bracketed integrand. We denote this expression by $\mathcal{K}$ and will find its minimum. Hence, $\mathcal{K} = \alpha T^{-1/2} f + \beta T g$ is a minimum with respect to T when

$$\mathcal{K}_{\min} = 3(\beta g)^{1/3} (\alpha f/2)^{2/3} . \qquad (8.9.10)$$

This results in

$$T = \left(\frac{\alpha f}{2\beta g}\right)^{2/3} . \qquad (8.9.11)$$

Now, substituting $\mathcal{K}_{\min}$ in the inequality (8.9.9), we get

$$\int_0^{z_i} \mathcal{K}_{\min} dz < \int_0^{z_i} \mathcal{K} dz < \alpha a + \beta b , \qquad (8.9.12)$$

so that

$$1.8\alpha^{2/3}\beta^{1/3} \int_0^{z_i} f^{2/3} g^{1/3} dz = 1.8\alpha^{2/3}\beta^{2/3}\psi(z_i) < \alpha a + b\beta , \qquad (8.9.13a)$$

or

$$\psi(z_i) < 0.55 \frac{(\alpha a + \beta b)}{\alpha^{2/3}\beta^{1/3}} . \qquad (8.9.13b)$$

The function ψ is defined by the integral presented above. Next, the ratio α/β is chosen so as to minimize the right-hand side of expression (8.9.13b). This procedure yields $\alpha/\beta = 2b/a$. Finally, having made this choice for α/β, we obtain from expression (8.9.13b) the result for ψ that

$$\psi(z_i) < a^{2/3} b^{1/3} , \qquad (8.9.14)$$

from which the upper limit on z_i is found. From this method,

$$z_i < 0.11 \left(\frac{W}{\Omega_0^5 h_0^{12}}\right)^{1/11} \quad 280(\Omega_0^5 h_0^{12})^{-1/11} . \tag{8.9.15}$$

The numerical value of z_i is given for a maximum energy release $W = 10^{15}$ ergs g^{-1}, which is about half of the nuclear energy of the gas. Another approach, based on estimates of the distortion, gives for $y < 0.03$

$$z_i < 500 \left(\frac{y^2}{\Omega_0^5 h_0^{12}}\right)^{1/9} < 250(\Omega_0^5 h_0^{12})^{-1/9} . \tag{8.9.16}$$

We see that for the very worst case, $\Omega_0 = 1$, $h_0 = 1$, we must have $z_i < 250$. Consequently, the gas is neutral during the period $1500 > z > 800$. Ideas about reionization taking place anywhere later than $z < 250$, after the recombination at $z = 1500$, inevitably follow from the absence of neutral hydrogen at $z \leq 3$. Less evident is the situation for $\Omega_0 < 1/10$ and $H_0 = 50$ km s^{-1} Mpc^{-1}. In this case the bound is not sufficient to allow one to be certain of the existence of a neutral-hydrogen period. Nevertheless, the formulae presented above do show how further measurements can help to resolve this problem.

8.10 The Interaction of Cosmic Rays with the Relic Radiation

Cosmic rays, which may propagate for an enormously long time in interstellar or intergalactic space, interact with the particles and fields that fill this space (Ginzburg and Syrovatsky 1963). After the discovery of RR, the interaction of the cosmic rays with the relic photons was discussed in many studies. A short account of the basic results is given below (for more detail, see Berezinsky 1970; Hayakawa 1973, 1974; Ozernoy, Rozental', and Prilutsky 1973).

A. The Interaction of the Radiation with Cosmic-Ray Electrons

The slowing-down of relativistic electrons ($E \gg m_e c^2$) is described by the formula

$$\frac{dE}{dt} = -\alpha \left(\frac{E}{m_e c^2}\right)^2 \epsilon_\gamma , \tag{8.10.1}$$

where $\alpha = 2.7 \times 10^{-14}$ cm^3 s^{-1} and where ϵ_γ is the total energy density in the electromagnetic field. This total energy density includes that in the magnetic fields $\epsilon_m = H^2/8\pi$, as well as the radiation energy density which consists of that from stars ϵ_{rs} and the RR energy ϵ_{RR}. The last equals 4×10^{-13} ergs cm^{-3} for a Planck spectrum at a temperature 2.7 K. If $\langle H \rangle \approx 10^{-6}$ gauss, then $\epsilon_m \approx 4 \times 10^{-14}$ ergs cm^{-3}, which is rather small compared with ϵ_{RR}; but if $\langle H \rangle \approx 10^{-5}$ gauss, then $\epsilon_m \approx 4 \times 10^{-12}$ ergs cm^{-3}, which is already larger than ϵ_{RR}. Inside a galaxy, $\epsilon_{rs} \approx 10^{-12}$ ergs cm^{-3}. Outside, ϵ_{rs} is of the order of 10^{-14} ergs cm^{-3}, which is much less than ϵ_{RR}. Here, it is essential that the photon energy be less than $m_e c^2$ in the electron's rest system. Then the Thomson scattering cross section applies

without large corrections from the Klein-Nishina-Tamm formula. This condition yields, for the RR, the limit $E_e < 10^{15}$ eV.

Equation (8.10.1) is easily integrated, yielding

$$E = \frac{E_0}{1 + (\alpha E_0 \epsilon_\gamma t)/(m_e c^2)^2} < \frac{(m_e c^2)^2}{\alpha t \epsilon_\gamma} . \tag{8.10.2}$$

For each pair ϵ_γ and t, an upper limit on the energy is given. Taking $t \sim 10^8$ years or 10^6 years and $\epsilon = 4 \times 10^{-13}$ ergs cm^{-3} or 4×10^{-11} ergs cm^{-3}, we obtain the upper limits shown in Table 9.

The energy removed from the electrons depends only on the total electromagnetic energy density ϵ_γ. But the form into which this energy transforms depends on the details of the electromagnetic field. For a static magnetic field $H \sim 10^{-5}$ to 10^{-6} gauss, synchrotron radiation is produced with a frequency (at the maximum of the spectrum)

$$\nu \approx 1.2 \times 10^6 \left(\frac{E}{m_e c^2}\right)^2 H_\perp \text{ Hz} . \tag{8.10.3}$$

Only the component $H_\perp$, the one perpendicular to the velocity of the electron, enters. The interaction of an electron with photons of energy $E_\gamma = h\nu$ gives photons with energy of the order of $E'_\gamma = (4E_\gamma/3)(E/m_e c^2)^2$.

It is necessary to indicate that all estimates depend heavily on the magnitude of the high-frequency part of the RR spectrum. There had been claims that the RR energy density was higher than that corresponding to a Planck spectrum at 2.7 K. It was pointed out that, in this case, the cosmic-ray electrons would produce a strong X-ray background. Later measurements of Blair *et al.*, however, have disproved these claims, and we are now reasonably certain that ϵ_{RR} is too low to account for the X-ray background.

B. The Interaction with Gamma Rays

Although the cross section is very small, there are photon-photon interactions. For sufficiently high energies, however, the creation of $e^\pm$ pairs can occur in such interactions. This process, the reverse of annihilation, has a large cross section of order 10^{-24} cm^2.

In order to calculate the required energy, we recall that $E^2 - c^2 p^2 = m_{\text{rest}}{}^2 c^4$, where m_{rest} is the rest mass. We apply this to the two photons with

TABLE 9

UPPER LIMITS ON THE ENERGY E OF ELECTRONS IN AN ELECTROMAGNETIC FIELD

ELECTROMAGNETIC FIELD ENERGY DENSITY (ergs cm^{-3})	LIFETIME (yr)	
	10^8	10^6
4×10^{-13}	2×10^{-2} ergs = 12 GeV	2 ergs = 1200 GeV
4×10^{-11}	2×10^{-4} ergs = 0.12 GeV	2×10^{-2} ergs = 12 GeV

energies E_1 and E_2 and momenta p_1 and p_2 in our system of coordinates. Here the index 1 will refer to the γ ray and the index 2 to the relic photon. Recall that $E_1 = c|p_1|$ and $E_2 = c|p_2|$. The total energy is $E_1 + E_2$, while the total momentum is $[|p_1|^2 + |p_2|^2 + 2|p_1||p_2| \cos\theta]^{1/2}$; we find $m_{\text{rest}}{}^2 c^4 \quad 2E_1 E_2 [1 - \cos\theta]$, where $\cos\theta = p_1 \cdot p_2 / |p_1||p_2|$.

To create a pair requires $m_{\text{rest}} c^2 > 2 m_e c^2$, so that $E_1 E_2 > m_e{}^2 c^4$ is necessary. Taking $E_2 = 4kT_0 \approx 10^{-3}$ eV, we obtain $E_1 \gtrsim 3 \times 10^{14}$ eV. The quenching of γ-rays more energetic than 300 TeV is thus very sharp. For the density of relic photons ($\approx 400\,\text{cm}^{-3}$) and for the pair-creation cross section ($\approx 10^{-24}\,\text{cm}^2$), we find that the lifetime of such an energetic γ-ray is $t \approx 10^{11}\,\text{s} \approx 3 \times 10^3\,\text{yr}$.

The experimental situation is not clear. Gamma-rays with energies $E > 100$ TeV are not observed, but lower-energy γ-rays are very few in number. It is therefore not known if the kink predicted for the region $\gtrsim 100$ TeV in the γ-ray spectrum actually exists.

C. *The Interaction of the Radiation with Heavy Particles*

More interesting is the interaction of the RR with heavy cosmic-ray particles such as the protons and other cosmic-ray nuclei. The direct electromagnetic interaction of such particles is, in general, about $(m_p/m_e)^2 \approx 4 \times 10^6$ times weaker than it is for electrons. Consequently, this interaction is usually negligibly small. But when the photon energy in the rest system of the heavy particles is sufficiently large, new processes can occur. These are (1) the creation of particle pairs, $\gamma + p \rightarrow e^+ + e^- + p$, with a threshold energy of $\Delta E \approx 1$ MeV; (2) photonuclear processes leading to the excitation or disintegration of a nucleus; and (3) the creation of pions, $\gamma + p \rightarrow p + \pi^0$ or $\gamma + p \rightarrow n + \pi^+$ with a threshold energy of $\Delta E \approx 140$ MeV (or higher for the creation of mesons heavier than pions).

Using as the laboratory energy of the photons, $E_\gamma \approx 10^{-3}$ eV, it is easy to calculate the proton energy required for any such process to proceed. The result is

$$E = \tfrac{1}{2} m_p c^2 \left(\frac{\Delta E}{E_\gamma}\right). \tag{8.10.4}$$

This gives $E \approx 10^9\, m_p c^2 \approx 10^{18}$ eV for the creation of a e^+e^- pair and $E \approx 10^{11}\, m_p c^2 \approx 10^{20}$ eV for the creation of a pion. The required energy grows in proportion to the mass of both the incident heavy particles and the final products.

A cosmic-ray particle with extremely high energy impinging on the Earth will give a "shower"—a large number of secondary particles—due to interactions in the atmosphere. These showers are studied by arrays of particle counters dispersed over areas of several square kilometers. By such means the existence of particles with energies up to 10^{22} eV has been established. The statistics are poor; we don't know for sure if these particles

are coming isotropically from all directions, or preferentially from a definite region of the sky.

The magnetic field in the Galaxy which bends trajectories of the "usual" cosmic rays ($E < 10^{15}$ eV) is not strong enough to perturb the motion of the 10^{22} eV projectiles.

If the most energetic particles do come isotropically, this would confirm their extragalactic origin with a correspondingly long mean travel time. This is surely incompatible with an overequilibrium ϵ_{RR} (Berezinsky and Zatsepin 1969), although not contradicting the Planckian value $\epsilon_{RR} = 4 \times 10^{-13}$.

The problem of the interaction of cosmic rays with relic neutrinos was discussed by Kacharov (1965). One can place a limit on the neutrino density in the Universe according to the gravitational influence of neutrinos on the expansion of the Universe (Zel'dovich and Smorodinsky 1961). From the ensuing estimate it follows that their (weak) interaction with the cosmic rays is insignificant. Let us return now to the main problem.

In the discussion of the interaction of cosmic rays presented above, it was assumed that cosmic rays are injected over a course of time of the order of the age τ_G of our Galaxy. If the cosmic-ray interaction time τ_{CR} is larger than τ_G, the interaction is not important. If, however, the interaction time is shorter, then the density of cosmic rays decreases as the ratio τ_{CR}/τ_G. If τ_{CR} depends on energy, then the characteristic value E_c for which $\tau_{CR}(E_c) = \tau_G$ must produce a kink in the cosmic-ray spectrum. It is not necessary to know the evolution of the temperature and the energy density of the RR background because τ_G is shorter than the cosmological time. Indeed, for charged particles in the cosmic rays contained by a galactic magnetic field, one cannot apply adiabatic losses of momentum and energy due to the Hubble expansion because the galaxies themselves do not expand.

Finally, a proposal was advanced not long ago according to which the most energetic cosmic rays are created nearby—for example, in the pulsar of the Crab Nebula. Then their age is small, and the contradictions related to the energy losses disappear.

Research into high-energy cosmic rays is unusually fascinating, but is very difficult because of the rarity of the events. It is necessary to collect a sufficient number of events in order to verify the energy and determine whether isotropy is destroyed.

Recently, hypothesized pregalactic cosmic rays were examined in connection with the conjecture that the observed γ-rays are remnants of the annihilation of matter and antimatter. In a completely homogeneous Universe with a homogeneously distributed excess in the number of baryons over that of antibaryons, annihilation is complete at $z \gg 10^{10}$ (see § 7.3). With the enormous density in this period, total equilibrium is established. Consequently, annihilation is a significant source of γ-rays only if the

process is delayed. A possible reason for a delay might be an inhomogeneous distribution of matter and antimatter. We postpone consideration of a detailed theory of inhomogeneous annihilation until the next part of the book, which is concerned with the inhomogeneous Universe. Here we shall consider only the problem of the propagation of γ-rays.

First of all, the thermal energy of the baryons and the antibaryons is low in comparison with $m_p c^2$ so that annihilation occurs for particles essentially at rest. The primary products consist mainly of mesons, but they are short-lived, so that only γ-rays and neutrinos survive. The energy in γ rays is about one-third of the total; the γ-ray spectrum has a maximum near 200 MeV. If the annihilation occurs at some redshift z, then the present energy of those γ-rays is z times less. On the basis of results of Vette *et al.* (1970), Stecker (1969) noticed a hump in the observed γ-ray spectrum around 1–6 MeV. As has been noted previously, Golenetsky *et al.* (1971) and Golenetsky and Mazets (1971) have shown that the results of Vette *et al.* (1970) are overestimated; there is no good evidence for the hump Stecker (1969) requires. He attributed this apparent hump to the decay of π^0 mesons produced at $z \approx 100$. Meanwhile, Arons and McCray (1969) have studied the propagation of energetic γ-rays with proper regard for their interaction with the plasma, and for the change in their energy and in the density of the plasma as the Universe expands.

When the energy of a γ-ray photon is greater than $m_e c^2$, the scattering cross section decreases in comparison to the Thomson cross section. Roughly,

$$\sigma = \sigma_T \frac{m_e c^2}{E} \ln \frac{E}{m_e c^2}, \quad E > m_e c^2. \tag{8.10.5}$$

An important fact concerning pure Thomson scattering is that the frequency does not change (when $h\nu \ll m_e c^2$). Consequently, the scattering does not change the soft part of the spectrum, integrated over angle. If the optical depth is large, then only the discrete sources are smeared out. On the other hand, if $h\nu \gg m_e c^2$, the energy losses of the photons are large. In such processes described below, the photon loses more than half its energy.

But, for a certain energy nucleus-mediated pair creation, $\gamma + (A, Z) \rightarrow e^+ + e^- + (A, Z)$, becomes more important than the scattering by electrons. For hydrogen, this threshold lies near 100 MeV. The result is that only those γ-rays with an energy $E_\gamma > 1$ MeV that are formed when $z < 1000$ can be detected today.

It is useful to compare this discussion with the indirect observations of annihilation due to energy release and to distortions in the spectrum. This method will work, within wide limits, right up to $z = 5 \times 10^4 \Omega_0^{-1/2}$. True, there will be no special features here; any process of energy release produces spectral distortions. Only observations of γ-rays and measurements of their spectrum and angular distribution will give direct evidence of annihilation. The already existing data are not convincing.

III GRAVITATIONAL INSTABILITY IN COSMOLOGY; GALAXY FORMATION

9 INTRODUCTION TO PART III

The question of the structure of the Universe, including a description of galaxies and clusters of galaxies, is one of the central questions in modern cosmology. The idealized picture of a strictly homogeneous and isotropic Universe set forth in Parts I and II undoubtedly reflects reality, but by no means all of reality—namely, this picture does not describe the origin of galaxies and clusters of galaxies.

The bases of the theory of the homogeneous Universe (including the theory of the hot Universe) were formulated more than twenty-five years ago. The last fifteen years have brought observations confirming the concept of the hot, homogeneous, isotropic model; they have hardly changed the basic theoretical viewpoint.

Concerning questions of the structure and inhomogeneity of the Universe, however, the situation is contradictory and the theory is in a state of confused development.

The high degree of RR isotropy suggests that perturbations were small at the epoch of hydrogen recombination, when $z \approx 1400$. On the other hand, a highly inhomogeneous matter distribution over galactic and cluster scales in the present epoch testifies to the growth of inhomogeneities from the recombination epoch to the present. Such growth of perturbations is naturally explained by the gravitational instability of a homogeneous distribution of matter. In Part III, the first chapters (chaps. 10, 11, and 12) are devoted to different aspects of the linear theory of instability: (1) the Newtonian approximation for an ordinary gas; (2) the theory of the hot Universe, where the interaction between ordinary matter and radiation is taken into account; and (3) using GTR, when the wavelength of the perturbations is comparable with the characteristic size of the Universe.

A mathematically exact theory can be developed only in the linear approximation, when the perturbations are small and therefore when each individual perturbation, each characteristic solution of the linear system of equations, evolves independently of the others.

Next, a short chapter is devoted to the question of random perturbations and of their description as a superposition of individual, characteristic solutions. Statistically specified functions are often employed in the theory of turbulence, in the theory of information transmission, and in the theory

of noise, but astrophysicists are perhaps insufficiently familiar with them. This subject is therefore explained in chapter 12.

A theoretical difficulty with this approach, however, is that the present structure of the Universe corresponds to large perturbations. The average density of matter in clusters of galaxies is several times larger than the average matter density in all of space, while in galaxies it is several orders of magnitude larger than the average. The linear theory of small perturbations is quite useless in describing such phenomena. Chapter 14, therefore, in which we set out a realistic description of the structure of the Universe, also contains a development of the theory of large-amplitude perturbations. In addition, we present there the theory describing the emergence of structure in the Universe, structure resulting from the evolution of small primordial inhomogeneities. This is the so-called theory of adiabatic perturbations. Also considered are other, alternative hypotheses for the formation of structure. These include the hypothesis of entropy perturbations and the hypothesis of a charge-symmetric Universe subdivided into regions of matter and antimatter.

The justification for this plethora of theories is that all modern theories describing the origin of galaxies have as a goal the calculation of the evolution of perturbations for given initial conditions. However, the choice of initial conditions is in large measure arbitrary. On what does the emergence of primordial perturbations therefore depend? What is their character? Only vague guesses exist concerning the answers to these questions (see chap. 15). Of course, choosing among initial perturbations of diverse character also leads to diverse theories for the origin of galaxies.

Only by comparing the consequences of theory with observations do we reach the conclusion that the theory of adiabatic perturbations is preferable to the others. It is necessary to emphasize, though, that the reliability of this conclusion is substantially less than the trustworthiness of assertions concerning the classical theory of evolution of the Universe as a whole, as set forth in Parts I and II of the book.

In chapter 16 we present discussions of the degree of homogeneity of the density distribution and of the isotropy of the expansion on a scale (greater than 100 Mpc) containing many clusters. In truth, the contrast is striking between the patchy distribution of material on a small scale and the smooth, regular distribution on a large scale.

One method of approach here involves studying RR. The results obtained up to now are in the form of inequalities—the perturbations are smaller than a definite limit; real perturbations have not been discovered yet. Estimates of the present-day distribution of matter with respect to the inhomogeneity on various scales and calculations of the evolution of inhomogeneity in the past, permit one, however roughly, to estimate the degree of initial inhomogeneity near the singularity. We thus outline

definite characteristics of the spectrum of initial perturbations and note a program of further research.

Chapter 17, devoted to gravitational waves, stands by itself. Though Weber's sensational experimental attempts to observe gravitational waves have not been confirmed, the experiments have certainly raised much interest. In particular, this is so in connection with their existence under cosmological conditions. Nevertheless, there remains a need for sober analysis of every wavelength band for gravitational waves and for an analysis of their possible significance in the evolution of the Universe. Since gravitational waves remain an example of a difficult-to-observe form of energy, however, it is necessary to make use of indirect methods in order to correctly evaluate their role.

10 GRAVITATIONAL INSTABILITY IN NEWTONIAN THEORY

10.1 JEANS'S THEORY

In this part of the book, we will examine the physical processes leading to the separation of the homogeneously expanding medium into separate globs.

The mathematical problem of the instability of a homogeneous distribution of matter was first posed and solved in the limit of small perturbations by James Jeans in 1902. This solution is set forth in detail in Jeans's well-known 1929 work. Regardless of a definite inconsistency, which will be noted and corrected in the subsequent account, Jeans's theory remains to this day of more than historical interest. Its value is underscored by the well rooted terminology such as the "Jeans length."

Jeans's theory is primarily an account of two factors: (1) gravity, which causes matter to accumulate in separate lumps or globs; and (2) pressure, which tends to smooth out inhomogeneities and to distribute the matter uniformly.

At first we will set forth the theory in the simplest possible form—in the form originally created by Jeans. Thus, recall the general equations for hydrodynamics and gravity describing an ideal gas in the Newtonian approximation

$$\begin{gathered} \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho \boldsymbol{u}) = 0\,; \\ \frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} + \frac{1}{\rho}\nabla P + \nabla\phi = 0\,. \\ \Delta\phi = \nabla\cdot\nabla\phi = 4\pi G\rho\,; \\ \frac{\partial S}{\partial t} + (\boldsymbol{u}\cdot\nabla)S = 0\,. \end{gathered} \tag{10.1.1}$$

Here ρ is the gas density, $\boldsymbol{u}$ is its velocity, S is the matter entropy, and ϕ is the gravitational potential. Assume, following Jeans, that the unperturbed state [denoted by the superscript (0)] is one in which the gas is at rest ($\boldsymbol{u} = 0$) and is uniformly distributed in space ($\rho - \rho^{(0)} = \text{const.}$, $S = S^{(0)} = \text{const.}$). Its pressure is everywhere the same ($P = P(\rho^{(0)}, S^{(0)}) = \text{const.}$). It is also tacitly assumed that the force of gravity disappears, $\nabla\phi = 0$, in the unbounded, uniform distribution, although this certainly contradicts Poisson's equation.

To obtain a perturbation solution, one ordinarily applies the method of decomposing an arbitrary perturbation according to some system of orthogonal functions—the Fourier method—seeking then a description of the time evolution of the constituent perturbations. In our case, the simplest procedure is to decompose the perturbation into a system of plane waves. We seek a perturbed solution in the form of one plane wave, with wave vector $\boldsymbol{k}$, superposed on the unperturbed solution. Thus, let us assume that

$$\begin{aligned}
\rho(x, t) &= \rho^{(0)}[1 + \delta(t) \exp(i\boldsymbol{k}\cdot\boldsymbol{x})] , \\
u(x, t) &= 0 + v(x, t) = w(t) \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) , \\
\phi(x, t) &= \phi^{(0)} + f(t) \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) , \\
S &= S^{(0)} + \sigma(t) \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) , \\
P &= P^{(0)} + (\partial P/\partial\rho)(\rho - \rho^{(0)}) + (\partial P/\partial S)(S - S^{(0)}) , \\
&= P^{(0)} + b^2\rho^{(0)}\delta \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) + h^2\sigma \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) .
\end{aligned} \tag{10.1.2}$$

In the expression for the pressure, the notation $(\partial P/\partial S) = h^2$ and $(\partial P/\partial\rho) = b^2$ is introduced, so that b is the adiabatic velocity of sound. Note that $\delta(t)$ is *not* the Dirac delta function.

Now insert these expressions into the equations of hydrodynamics. As is appropriate in perturbation theory, we consider only terms linear in the quantities δ, w, f and σ. The zero-order terms (not containing δ, w, f, or σ) describe the unperturbed solution. For these it is assumed that the equations are satisfied identically. The second-order terms (quadratic in the products of the small quantities, such as δ^2 and $\delta\sigma$) are neglected. This leads to the following system of linear, homogeneous equations (where $k = |\boldsymbol{k}|$):

$$\begin{gathered}
\frac{d\delta}{dt} + i\boldsymbol{k}\cdot\boldsymbol{w} = 0 ; \\
\frac{dw}{dt} + ikf + ikb^2\delta + ikh^2[\rho^{(0)}]^{-1}\sigma = 0 ; \\
k^2 f = -4\pi G\rho^{(0)}\delta ; \\
\frac{d\sigma}{dt} = 0 .
\end{gathered} \tag{10.1.3}$$

In Jeans's theory, the unperturbed solution does not depend on time. Consequently, the coefficients $\rho^{(0)}$, b^2, and $\boldsymbol{k}$ are constant in the perturbation equations (10.1.3). Time enters only through the derivatives. In this case, the general theory predicts a definite—in particular, exponential—character for the time dependence of the perturbations:

$$\delta = \delta_0 e^{\omega t} , \quad w = w_0 e^{\omega t} , \quad f = f_0 e^{\omega t} , \quad \sigma = \sigma_0 e^{\omega t} . \tag{10.1.4}$$

The discussion will run approximately as follows. Denote the solution of the system by $F(t)$, where F represents all of the quantities, δ, w, f, and σ in the given relations. By virtue of the linearity and homogeneity of the

equations, the quantity $AF(t)$, where A is a constant factor, is also a solution. On the other hand, because the coefficients in the equations are constant, the solution permits a shift in time: $F(t + \tau)$ is also a solution. It is assumed that there exist solutions ("characteristic" solutions) for which both groups of transformations give one and the same result, i.e., for which $AF(t) = F(t + \tau)$. Then, for each value of τ, there is at least one corresponding value of A. Such a functional relation holds if $F = F_0 e^{\omega t}$ and $A = e^{\omega t}$.

The problem for the theory then consists of the determination of the quantity ω and of the relations among the various parameters (δ, w, f, and σ) in the characteristic solutions. The absolute amplitude of the proper solution depends on the initial conditions.

Let us now return to our problem. Inserting the expressions (10.1.4) into the system (10.1.3), we obtain

$$\begin{aligned} \omega\delta_0 &= -\,i\boldsymbol{k}\cdot\boldsymbol{w}_0\,; \\ \omega\boldsymbol{w}_0 &= -i\boldsymbol{k}f_0 - i\boldsymbol{k}b^2\delta_0 - i\boldsymbol{k}[\rho^{(0)}]^{-1}h^2\sigma_0\,; \\ -k^2 f_0 &= 4\pi G\rho^{(0)}\delta_0\,; \\ \omega\sigma_0 &= 0\,. \end{aligned} \tag{10.1.5}$$

A. Adiabatic Perturbations

Of most interest is the time-dependent solution with $\omega \neq 0$. When $\sigma_0 = 0$ and the vector $\boldsymbol{w}$ is parallel to the wave vector $\boldsymbol{k}$, we introduce the notation

$$\boldsymbol{w}_0 = w_0\boldsymbol{k}/k\,,$$

and obtain a system of scalar equations for the time-dependent perturbations with $\omega \neq 0$. These equations are

$$\begin{aligned} \omega\delta_0 &= -ikw_0\,; \\ \omega w_0 &= -ikf_0 - ikb^2\delta_0\,; \\ -k^2 f_0 &= 4\pi G\rho^{(0)}\delta_0\,. \end{aligned} \tag{10.1.6}$$

System (10.1.6) has a nontrivial solution if the condition

$$\omega = \pm(4\pi G\rho^{(0)} - b^2k^2)^{1/2} \tag{10.1.7}$$

is satisfied. It is clear that the properties of the solution depend critically on the sign of the expression above whose square root is to be taken. For specific matter properties ($\rho^{(0)}$ and b^2), there exists a critical value k_J determined by the condition $\omega = 0$:

$$\begin{aligned} &4\pi G\rho^{(0)} - b^2k_J^2 = 0\,; \\ &k_J = \frac{1}{b}[4\pi G\rho^{(0)}]^{1/2}\,, \qquad \lambda_J = \frac{2\pi}{k_J} = b\left[\frac{\pi}{G\rho^{(0)}}\right]^{1/2}. \end{aligned} \tag{10.1.8}$$

Here λ_J is the critical wavelength (the Jeans length) corresponding to the critical value k_J (the index J is short for Jeans).

We will now examine separately the consequences of the theory of gravitational instability for the cases $\lambda > \lambda_J$ ($k < k_J$) and $\lambda < \lambda_J$ ($k > k_J$).

For $\lambda > \lambda_J$, the quantity ω, determined through expression (10.1.7), is real; it can have a positive or a negative value. The existence of a solution with positive ω, the solution growing as $e^{\omega t}$, signifies that a homogeneous distribution of matter is gravitationally unstable. From the equations it follows that for real ω, the ratio of δ_0 to w_0 contains the imaginary unit i. As is well known, when one uses complex quantities in calculations relating to real variables such as density, pressure, and velocity, what is meant is that the equations are satisfied for the real parts of the complex variables. We will assume that δ_0 is real, so we write the growing solution in the real form

$$\begin{aligned} \rho &= \operatorname{Re} \rho^{(0)}(1 + \delta_0 e^{\omega t + i\boldsymbol{k}\cdot\boldsymbol{x}}) = \rho^{(0)}(1 + \delta_0 e^{\omega t} \cos \boldsymbol{k}\cdot\boldsymbol{x}) ; \\ \boldsymbol{v} &= \operatorname{Re}\left(\frac{\boldsymbol{k}}{k} e^{i\boldsymbol{k}\cdot\boldsymbol{x}} i \frac{\omega}{k} \delta_0 e^{\omega t}\right) = -\frac{\delta_0 \boldsymbol{k}}{k^2} \omega e^{\omega t} \sin \boldsymbol{k}\cdot\boldsymbol{x} . \end{aligned} \tag{10.1.9}$$

Thus, the imaginary units in the ratio of the perturbed density and the velocity signify that these quantities are shifted in phase with respect to each other. If the density is at a minimum or a maximum, the velocity is zero; when the magnitude of the velocity is at a maximum, the density perturbation is zero. For $\lambda \to \infty$ (the wave vector $k \to 0$), the frequency $\omega \to (4\pi G\rho^{(0)})^{1/2}$. The perturbation grows by a factor of e in the time $\tau = 1/\omega = (4\pi G\rho^{(0)})^{1/2}$, which, in order of magnitude, coincides with the cosmological time. The latter is the time it takes, in the Friedmann model, for the density to change from $\rho = \infty$ to $\rho = \rho^{(0)}$; this follows from the relation $\rho^{-1} = (6\pi Gt^2)$ or $t = (6\pi G\rho)^{-1/2}$. This coincidence, moreover, does not occur by chance.

For the second case, $\lambda < \lambda_J$ or $k > k_J$, it is better to rewrite formula (10.1.7) in the form

$$\omega = \pm i(b^2k^2 - 4\pi G\rho^{(0)})^{1/2} . \tag{10.1.10}$$

Then, if the frequency ω is imaginary, the ratio of the density perturbation δ_0 to the velocity is real. This ratio then denotes that the perturbed density and the velocity have the same phase. A maximum in the velocity coincides with a maximum or a minimum in the density. At the same time, the characteristic solution has the properties of a traveling wave. At each point the velocity in the propagation direction is proportional to the perturbation in the density. The amplitude, meanwhile, does not depend on time. This characteristic solution is

$$\begin{aligned} \rho &= \rho^{(0)}[1 + \delta_0 \cos (\boldsymbol{k}\cdot\boldsymbol{x} - |\omega| t)] ; \\ \boldsymbol{v} &= \frac{\boldsymbol{k}|\omega|}{k^2} \delta_0 \cos (\boldsymbol{k}\cdot\boldsymbol{x} - |\omega| t) ; \\ |\omega| &= (k^2b^2 - 4\pi G\rho^{(0)})^{1/2} . \end{aligned} \tag{10.1.11}$$

Let us examine the limiting case of very short wavelengths, $k^2 \gg k_J^2 = 4\pi G\rho^{(0)}/b^2$. In expression (10.1.10), one can then neglect the quantity $4\pi G\rho^{(0)}$ in comparison to k^2b^2. Solution (10.1.11) then becomes a description of a simple sound wave propagating at the velocity of sound. Given the specific initial conditions, $t = 0$, $u = 0$ everywhere, the solution takes the form of standing waves:

$$\rho = \rho^{(0)}[1 + \delta^{(0)} \cos(\boldsymbol{k}\cdot\boldsymbol{x}) \cos(\omega t)] ;$$

$$v = \boldsymbol{k}\left(\frac{\omega}{k^2}\right)\delta^{(0)} \sin(\boldsymbol{k}\cdot\boldsymbol{x}) \sin(\omega t) .$$

It will be shown below that just such a solution is obtained in the cosmological problem.

Thus, there are two, distinct, limiting cases. For a long-wavelength perturbation, gravity plays the primary role. Gravity is responsible for the instability of a uniform matter distribution and for the existence of exponentially growing density perturbations. In this case, one can neglect pressure. For short-wavelength perturbations, however, pressure plays the dominant role. A density perturbation is accompanied by a pressure perturbation, which leads to the propagation of an acoustic wave of constant amplitude. Given this situation, one can thereafter neglect the role of gravity. In the next approximation, gravity will change the sound velocity, but the amplitude of the perturbations remains constant and no instability appears.

The boundary beyond which gravitational instability occurs is determined by the condition $\lambda = \lambda_J$, i.e., this boundary is at the Jeans length. One often speaks also of the "Jeans mass"—in the volume $(\lambda_J/2)^3$—

$$M_J \equiv (\lambda_J/2)^3\rho^{(0)} . \qquad (10.1.12)$$

This is the smallest mass for which the pressure cannot prevent growth in the density under the action of gravity. It is therefore assumed that in the course of time, a homogeneous distribution of matter will separate into discrete objects, each with a mass not less than M_J. The particular masses (but not less than M_J) into which the medium fragments depends on the spectrum of small initial perturbations, about which a more detailed discussion will be given later. Originally, the smaller the scale of a perturbation, the greater its amplitude. Therefore, the medium fragments at the smallest masses for which the pressure does not interfere with the condensation process, i.e., at a mass $\sim M_J$.

It is convenient to express M_J and λ_J in terms of units common in astronomy:

$$\lambda_J = 27\left(\frac{T}{\mu}\right)^{1/2}[\rho_{-24}^{(0)}]^{1/2} \text{ pc} ,$$
$$M_J = 33\left(\frac{T}{\mu}\right)^{3/2}[\rho_{-24}^{(0)}]^{-1/2} M_\odot . \qquad (10.1.13)$$

Here, μ is the mean molecular weight ($\mu = 1$ for neutral hydrogen, $\mu = \frac{1}{2}$ for ionized hydrogen, etc.), the temperature T is in kelvins, and $\rho_{-24}^{(0)}$ is the unperturbed density in units of 10^{-24} g cm^{-3}. For pure hydrogen, $n_H \approx 0.6\rho_{-24}$; also, 1 pc $= 3.1 \times 10^{18}$ cm and $M_\odot = 2 \times 10^{33}$ g.

In the form in which it was set forth above, the Jeans theory is formally incorrect. This is so because the unperturbed, homogeneous matter distribution is assumed to be stationary. In reality, inasmuch as the constant density $\rho^{(0)}$ corresponds to a varying gravitational potential ϕ, the unperturbed solution must be nonstationary. One cannot cure this "disease" by taking for the unperturbed solution a static solution in which gravitation is equalized by a pressure gradient. For, as is shown in Volume 1, such a configuration must have a finite mass. Also, the resultant body must have a size on the order of λ_J, so that the theory of Jeans is not applicable to it. On the other hand, a homogeneous unbounded distribution of matter must be nonstationary, i.e., it must possess a time-dependent density and must undergo general expansion (or contraction).

At first glance, a nonstationary solution differs sharply from the stationary one employed in Jeans's theory. But as we shall see, all the results of the exact theory are quite close to Jeans's results. The formally incorrect Jeans theory is a good approximation, and it does help one to understand the more complicated exact theory.

How ought one to generalize the Jeans theory?

1. By taking account of the time dependence $\rho^{(0)} = \rho^{(0)}(t)$, with $\rho^{(0)}(t)$ determined by the cosmological solution.

2. By taking account, also, of the consequent time dependence of the pressure and sound velocity: $P^{(0)} = P(\rho^{(0)}) = P[\rho^{(0)}(t)]$.

3. By taking account of the growth of wavelengths during the expansion. They increase in the same way as does the distance between every pair of points: $\lambda \sim a(t)$, $k \sim a^{-1}(t)$, $\rho^{(0)} \sim a^{-3}(t)$.

4. Finally, by substituting these improved forms of ρ, k, and b into equations (10.1.7) and (10.1.10). This leads to an expression for $\omega = \omega(t)$. The evident generalization of the expression $e^{\omega t}$ is then $\exp[\int \omega dt]$. This corresponds to the assumption that the perturbations satisfy a differential equation of the form $d\delta/dt = \omega(t)\delta$ in the nonstationary case with $\omega(t)$ taken from stationary theory. In § 10.3 the results obtained with the help of this generalization of Jeans's results are compared with the exact solution of the problem of the gravitational instability of homogeneous matter. Good agreement between them shows that the "incorrect" theory is, in essence, right.

B. Entropy and Vortex Perturbations

Let us then return to Jeans formulation, with a stationary background. Consider the solutions corresponding to zero frequency, i.e., to those

perturbations which do not depend on time at all. The condition $\omega\sigma = 0$ for an entropy perturbation permits a solution with $\sigma \neq 0$. It is evident that in the absence of thermal conductivity, an initial nonuniformity in the entropy distribution is preserved in the course of time. The complete solution with $\omega = 0$ also requires mechanical equilibrium. We thus obtain from the equation of motion

$$\delta_0 = \frac{-k^2\sigma}{\rho^{(0)}b^2 - 4\pi G[k^{-1}\rho^{(0)}]^2}. \tag{10.1.14}$$

For large k, i.e., for small wavelength, one can neglect the second term in the denominator. This term represents the forces of gravity. Under this circumstance, the connection between δ_0 and σ corresponds to the simple condition of the constancy of the pressure. When accounting for the force of gravity, however, the equilibrium condition changes. One can therefore convince oneself that in the problem with finite, nonzero values of δ_0 and σ at $t = 0$, the value of δ_0 at all times $t \neq \infty$ remains finite. For $k = k_J$, however, this quantity consists of the difference of two infinite quantities.

Another type of solution with $\omega = 0$ is found if the vector v is taken perpendicular to the wave vector, so that $v \cdot k = 0$. From the equation $\omega = 0$, it then follows that one can also choose $\sigma = 0$ (the case $\sigma \neq 0$ was singled out above) with $\delta_0 = 0$ and $f = 0$. When $v \perp k$, the divergence of the velocity is evidently equal to zero. This is in complete correspondence with the constancy of the density for such motion. However, the curl of the velocity,

$$\nabla \times v = i[w \times k], \tag{10.1.15}$$

is different from zero. Such motion is called vortex motion. For a given direction k, there are two directions perpendicular to k and to each other. In other words, there can exist two linearly independent vortex motions.

Let us count up the total number of types of perturbations for a given k. There are two longitudinal perturbations, one with $\omega = (4\pi G\rho^{(0)} - b^2k^2)^{1/2}$ and one with $\omega = -(4\pi G\rho^{(0)} - b^2k^2)^{1/2}$, one entropy perturbation, and two vortex perturbations, for a total of five.

The initial state of the medium is characterized by specified values of the density, entropy, and velocity. The velocity is a vector quantity, described by three components. Consequently, the initial state is characterized by specifying five quantities in all. This just allows one to determine unambiguously the amplitude of the five independent types of perturbations.

Up to now, only perturbations with a prescribed periodic dependence on the spatial coordinates ($e^{ik\cdot x}$) have been examined. Perturbations with an arbitrary dependence on the coordinates must at first be decomposed—in a Fourier integral—into periodic components. This decomposition will be discussed later in chapter 13. But now let us continue the examination of those particular waves with periodic dependence on the coordinates.

10.2 The Instability of Expanding, Homogeneous Matter

The instability of expanding, homogeneous matter was demonstrated in the Newtonian approximation by Bonnor (1957). He used as an unperturbed solution the nonstationary, isotropic cosmological models examined in the first part of this book. Recall the solution describing the evolution of isotropically expanding, homogeneous matter. The exact solution of the unperturbed system of equations (10.1.1) has the form

$$u = H(t)x\ ; \quad \phi^{(0)} = \tfrac{2}{3}\pi G\rho^{(0)}(t)x^2\ ; \quad \rho = \rho^{(0)}(t)\ ; \quad P = P^{(0)}(t)\ . \tag{10.2.1}$$

Here x is the radius vector, $x^2 = x_1^2 + x_2^2 + x_3^2$, while $H(t)$ and $\rho^{(0)}(t)$ satisfy the equations

$$\frac{d\rho^{(0)}(t)}{dt} = -3H\rho^{(0)}\ ; \quad \frac{dH}{dt} + H^2 = -\frac{4\pi G}{3}\rho^{(0)}\ . \tag{10.2.2}$$

All of the equations, including the Poisson equation, are satisfied by the solution (10.2.1). It is true, however, that the price for this is an infinite potential and an infinite velocity at spatial infinity: $u \to \infty$ for $x \to \infty$.

We seek the perturbations of the form

$$\frac{\rho - \rho^{(0)}}{\rho^{(0)}} = \delta(t)e^{ik\cdot x} = \delta(t)e^{i\varkappa\cdot x/a(t)}\ , \qquad \varkappa = \text{const.}\ ; \quad k = \varkappa/a(t)\ . \tag{10.2.3}$$

In agreement with the remarks at the end of § 10.1, the wavelength of a perturbation increases as the Universe expands. For perturbations of the velocity and the gravitational potential, we write

$$v = w(t)\frac{k}{k}e^{ik\cdot x}\ ; \quad \phi = \phi^{(0)} + f(t)e^{ik\cdot x}\ . \tag{10.2.4}$$

The velocity perturbations are assumed to be derivable from a potential (irrotational), i.e., $v \parallel k$. The unperturbed motion of a particle is described by the equation

$$\frac{dx}{dt} = u(x, t) = H(t)x\ ,$$

the general solution of which takes the form

$$x = a(t)q\ ,$$

where $a(t) = \exp\left[\int H(t)dt\right]$.

In the fully relativistic theory, $a(t)$ determines the radius of the Universe. In the nonrelativistic Newtonian theory, $a(t)$ is simply the scale factor specifying the changes in distance between every pair of particles. In Newtonian theory, the absolute value of $a(t)$ is unimportant; only the ratio $a(t_1)/a(t_2)$ has meaning. The quantity q is the Lagrangian coordinate, which

is constant (independent of time) for a given particle, while the vector $\boldsymbol{\kappa}$ is the wave vector expressed in Lagrangian coordinates.

We assumed above that the density perturbations take the form

$$\delta = [\rho(\boldsymbol{x}, t) - \rho^{(0)}(t)]/\rho^{(0)}(t) = \delta(t)e^{i\boldsymbol{\kappa}\cdot\boldsymbol{q}} ,$$

with analogous expressions holding for the velocity and the potential. The validity of that assumption is indirectly tested by substituting the quantities δ, w, and f into the fundamental equations. Taking into account that $\boldsymbol{k} = \boldsymbol{k}(t)$ and that $\boldsymbol{\kappa}$ = const., one can bring the final equations for the quantities δ, w, and f into the form (the factor $e^{i\boldsymbol{\kappa}\cdot\boldsymbol{q}}$ cancels out)

$$\begin{aligned} \frac{d\delta}{dt} &= -ikw ; \\ \frac{dw}{dt} + Hw &= i\frac{\delta}{k}(4\pi G\rho^{(0)} - k^2b^2) ; \\ f &= -\frac{4\pi G\rho^{(0)}\delta}{k^2} . \end{aligned} \tag{10.2.5}$$

This is a system of two first-order equations with time-dependent coefficients $k(t)$, $H(t)$, $\rho^{(0)}(t)$, and $b(t)$. (We will omit the superscript on $\rho^{(0)}$ in the subsequent development.) These coefficients are determined by the unperturbed solution and by the equation of state of the gas. For $H \equiv 0$ and ρ, k, b = const., the system (10.2.5) passes into the Jeans equation. In general though, it is easy to obtain from equations (10.2.5) the single second-order equation

$$\frac{d^2\delta}{dt^2} + 2H\frac{d\delta}{dt} - (4\pi G\rho - b^2k^2)\delta = 0 . \tag{10.2.6}$$

As studies have shown, the general solution to this equation is rather cumbersome. We therefore will begin with the simplest cases.

Equation (10.2.6) simplifies greatly if it is assumed that $P \sim \rho^{4/3}$, so that the velocity of sound satisfies $b = (dP/d\rho)^{1/2} \sim \rho^{1/6}$. Notice too that $\rho \sim a^{-3}$ and $k^2 \sim a^{-2}$. Consequently, both terms in the parentheses in equation (10.2.6) are proportional to a^{-3}; i.e., their ratio is constant. In this case, for a given comoving volume (for a given mass) the ratio of the internal energy and the gravitational energy remains constant during the expansion. These same properties are also manifested by the perturbations.

Let us now make a second simplifying assumption. Consider the case of the flat model (Einstein–de Sitter model) for which the unperturbed motion is described by the formulae

$$\begin{gathered} \rho = (6\pi Gt^2)^{-1} ; \quad a \propto t^{2/3} ; \quad k \propto t^{-2/3} ; \\ H = \tfrac{2}{3}t^{-1} ; \quad b \propto t^{-1/3} . \end{gathered}$$

Then equation (10.2.6) reduces to the form

$$t^2\frac{d^2\delta}{dt^2} + \tfrac{4}{3}t\frac{d\delta}{dt} - (\tfrac{2}{3} - b_1^2k_1^2)\delta = 0 . \tag{10.2.7}$$

This equation has the power law solutions

$$\delta = \delta_1 t^n ; \quad w = w_1 t^{n-1/3} , \quad n = -\tfrac{1}{6} \pm (\tfrac{25}{36} - b_1^2 k_1^2)^{1/2} . \quad (10.2.8)$$

It is evident that the solutions depend critically on the sign of the expression above whose square root is to be taken.

It is instructive to compare this result with that of the Jeans theory. From equations (10.2.8), we have that the argument of the square root in the expression for n goes to zero (at which point $n = -\frac{1}{6}$) when

$$\lambda_{cr} = \frac{2\pi}{k_{cr}} t^{2/3} = \tfrac{6}{5} 2\pi b_1 t^{2/3} = \tfrac{6}{5} 2\pi b t . \quad (10.2.9)$$

For comparison purposes we recall that the critical wavelength in the Jeans theory is defined by the formula $\lambda_J = 2\pi b(4\pi G\rho)^{-1/2}$. Then, substituting $\rho = (6\pi G t^2)^{-1}$ into this expression, we obtain

$$\lambda_J = (\tfrac{3}{2})^{1/2} 2\pi b t . \quad (10.2.10)$$

In an expanding Universe, the critical wavelength is proportional and, in order of magnitude, equal to the distance bt or, more precisely,

$$\int_0^t b dt ,$$

that sound travels during the cosmological expansion. This result will be widely used below. All of the differences between the exact value of the critical Jeans wavelength and the approximate value obtained from the Jeans theory itself, is due to the difference between the factors $6/5 = 1.2$ and $\sqrt{(3/2)} = 1.225$. Other conclusions of the Jeans theory are also nearly the same as those of the exact theory.

For $k \ll k_{cr}$, we obtain $n_1 = \frac{2}{3}$ and $n_2 = -1$. Generalizing Jeans's results (as indicated in § 10.1), we then obtain

$$\delta = \exp [\pm \int (4\pi G\rho)^{1/2} dt] = \exp [\pm (2/3)^{1/2} \ln t] = t^{\pm\sqrt{(2/3)}} .$$

In Bonnor's exact theory, the other limiting case, $k \gg k_{cr}$, corresponds to the sound waves

$$\delta = t^{-1/6 \pm i b_1 k_1} e^{i \mathbf{k} \cdot \mathbf{x}} = t^{-1/6} \exp [i \mathbf{k} \cdot \mathbf{x} \pm i b_1 k_1 \ln t] . \quad (10.2.11)$$

The phase velocity coincides with the instantaneous sound velocity:

$$\frac{dx}{dt} = \frac{b_1 k_1}{|k_1| t^{-2/3}} \frac{d \ln t}{dt} = b_1 t^{-1/3} = b . \quad (10.2.12)$$

The appearance in formula (10.2.11) of the factor $t^{-1/6}$ is easy to explain with the help of the theory of adiabatic invariants. In Paradoksov's amusing article (1966) "How Quantum Theory Helps One to Understand Classical Mechanics," it is explained that if $E = Nh\nu$, where N is the number of quanta in a given wave or in a given excited state of an oscillator, then

N = const. for a slow change in the parameters. In agreement with this theory, sound-wave energy changes in proportion to the frequency of the wave: $E \propto \nu$. The frequency satisfies $\nu = b/\lambda \sim b_1 t^{-1/3} k_1 t^{-2/3} \sim t^{-1}$. Meanwhile, the sound energy in the comoving volume V is $E \sim \rho v^2 V \sim \rho V (b\delta)^2 \sim (b\delta)^2 \sim t^{-2/3}\delta^2$. Therefore, $\delta^2 \sim t^{-1} t^{2/3} \sim t^{-1/3}$, so $\delta \sim t^{-1/6}$.

If, upon generalizing the Jeans theory, one takes account of this factor also for aperiodic waves, one then obtains the law $t^{-1/6 \pm \sqrt{(2/3)}}$ which agrees well with the exact solutions $t^{2/3}$ and t^{-1}. Instead of $\frac{2}{3} = 0.67$ and -1, one obtains approximately $[-\frac{1}{6} + (2/3)^{1/2} \approx 0.65; -\frac{1}{6} - (2/3)^{1/2} \approx -0.98]$. In Bonnor's (1957) work the time evolution of spherical waves, not flat plane waves, was examined. It is evident that this cannot change the results of the linear theory since one can construct a spherical wave with the help of a suitable combination of plane waves.

Notice finally that it is also necessary to include entropy and vortex perturbations in the general classification of perturbations of the expanding Universe. This requirement is as valid here as it is for the Jeans theory of § 10.1.

Entropy perturbations do not depend on time (only the wavelength grows in proportion to the general expansion). They satisfy

$$\sigma(\boldsymbol{x}, t) = \sigma_0 e^{i\boldsymbol{k}\cdot\boldsymbol{x}}; \quad \boldsymbol{k}(t) = \boldsymbol{\kappa}/a(t); \qquad \sigma_0 = \text{const.}; \quad \boldsymbol{\kappa} = \text{const}. \tag{10.2.13}$$

The expressions for the corresponding density perturbations in the general case of an arbitrary equation of state are rather complicated. In principle, an initial entropy perturbation can give rise to growing density perturbations. We will return to this question later.

Vortex perturbations in the expanding Universe constitute another form of velocity perturbation, and they are directed perpendicular to the wave vector, just as was the case in § 10.1:

$$\boldsymbol{v}(\boldsymbol{x}, t) = \boldsymbol{w}(t) e^{i\boldsymbol{k}\cdot\boldsymbol{x}}; \quad \boldsymbol{w}\cdot\boldsymbol{k} = 0. \tag{10.2.14}$$

At the same time, the perturbations of all other quantities—the density, the potential, and the entropy—are identically equal to zero.

For the amplitude of the vortex velocity, the equation

$$\frac{d\boldsymbol{w}}{dt} + H(t)\boldsymbol{w} = 0$$

is obtained. Comparing it with the expansion law

$$\frac{da}{dt} - H(t)a = 0,$$

we find the result

$$\boldsymbol{w}(t) = \frac{\boldsymbol{w}_0}{a(t)}, \tag{10.2.15}$$

where $\boldsymbol{w}_0$ is a constant vector perpendicular to the wave vector $\boldsymbol{k}$. The magnitude and direction of $\boldsymbol{w}_0$ are determined by the initial conditions. This result is valid for any law governing the change in $a(t)$ and, in particular, in the case corresponding to a closed ($\Omega_0 < 1$) or an open ($\Omega_0 > 1$) model. For an isolated wave, i.e., in the absence of perturbations of another type and of perturbations with different wave vectors, the result is also valid for finite, large-amplitude values of $\boldsymbol{w}_0$.

In the following section we will return to the analysis of increasing and decreasing density perturbations, which are accompanied by longitudinal matter motion, i.e., $\boldsymbol{v} \parallel \boldsymbol{k}$. The general case of an arbitrary equation of state $P = P(\rho)$ and of a nonflat Universe leads to complicated calculations and is not considered here (see Hunter 1962; Savedoff and Vila 1962). But the case of small $\boldsymbol{k}$ will be studied in detail also for a nonflat universe. When $\boldsymbol{k}$ is small, the equation of state of the matter does not enter into the equations of the problem; thus one complicating factor is eliminated.

10.3. Large-Scale Perturbations; the Self-similar Solution

The results obtained above show that pressure does not influence the development of large-scale perturbations. This fact is quite important and deserves a more detailed discussion.

Within the realm of the Fourier method, the equations for density perturbations δ and for the product $i\boldsymbol{k}\cdot\boldsymbol{w} \equiv \mathfrak{D}$ do not depend on k in the limit $k \to 0$. Consequently, perturbations with any wavelength satisfying $k \ll k_J$ develop with time according to one and the same law. In this case, therefore, the use of a Fourier decomposition is unnecessary. More precisely, for each wave there are two solutions—for the density perturbation and the longitudinal velocity—one growing and one decaying ($\delta \sim t^{2/3}$ and $\delta \sim t^{-1}$ in the simplest case—the flat model). One can therefore foresee that the solution of the equations of motion for the matter, with account taken of gravity but not of pressure, can be presented in the form

$$\begin{aligned} \rho(x, t) &= \rho^{(0)}(t)[1 + \delta_i(t)\psi_i(q) + \delta_d(t)\psi_d(q)] ; \\ \mathfrak{D}(\boldsymbol{x}, t) &= -\psi_i(q)\frac{d\delta_i}{dt} - \psi_d(q)\frac{d\delta_d}{dt} . \end{aligned} \tag{10.3.1}$$

For one wave, the product $i\boldsymbol{k}\cdot\boldsymbol{w}$ equals the divergence of the velocity. But let us generalize to the case of an arbitrary motion with $\mathfrak{D} = \nabla\cdot\boldsymbol{w}$. Two separate solutions appear in a factored form, i.e., they are written as the product of a function of time and a function of the coordinates. One solution with the subscript i (for "increasing"), grows, while the other, with the index d (for "decreasing"), decays. Since the perturbed region participates in the general expansion of the Universe, the factorization is achieved only by the use of functions depending on $\boldsymbol{q}$, i.e., on Lagrangian coordinates.

Let us digress from a consideration of vortex perturbations, whose time

dependence differs from that of $\delta_i(t)$ and $\delta_d(t)$. This is done by considering $\mathfrak{D}$, the divergence of the velocity, (not the velocity v itself) since the divergence of the vortex part is identically zero. The two functions of time δ_i and δ_d represent the solution of one second-order equation, namely,

$$\frac{d^2\delta}{dt^2} + 2H\frac{d\delta}{dt} - 4\pi G\rho\delta = 0 . \tag{10.3.2}$$

This equation coincides with equation (10.2.6), obtained by a Fourier decomposition, if, in the latter, one sets $k = 0$. This is not surprising since, evidently, $e^{i\boldsymbol{k}\cdot\boldsymbol{x}} = e^{i\boldsymbol{\varkappa}\cdot\boldsymbol{q}}$ is a particular case of the function $\psi(\boldsymbol{q})$. Equation (10.3.2) possesses two linearly independent solutions, δ_i and δ_d. For example, they might be the $t^{2/3}$ and t^{-1} mentioned earlier in connection with the simplest case—the flat model. Therefore, if $\delta(\boldsymbol{x}, t)$ and $\mathfrak{D}(\boldsymbol{x}, t)$ are specified at $t = t_0$, then, by introducing $\boldsymbol{q}$ in place of $\boldsymbol{x}$, one can obtain the two equations:

$$\begin{aligned} \delta_i(t_0)\psi_i(\boldsymbol{q}) + \delta_d(t_0)\psi_d(\boldsymbol{q}) &= \delta(t_0, \boldsymbol{q}) , \\ \frac{d\delta_i(t_0)}{dt}\psi_i(\boldsymbol{q}) + \frac{d\delta_d(t_0)}{dt}\psi_d(\boldsymbol{q}) &= -\mathfrak{D}(t_0, \boldsymbol{q}) . \end{aligned} \tag{10.3.3}$$

Solving these equations with respect to $\psi_i(\boldsymbol{q})$ and $\psi_d(\boldsymbol{q})$, we find

$$\begin{aligned} \psi_i(\boldsymbol{q}) &= \frac{\delta(t_0, \boldsymbol{q})d\delta_d(t_0)/dt + \mathfrak{D}(t_0, \boldsymbol{q})\delta_d(t_0)}{\delta_i(t_0)d\delta_d(t_0)/dt - d\delta_i(t_0)/dt\delta_d(t_0)} \\ \psi_d(\boldsymbol{q}) &= -\frac{\delta(t_0, \boldsymbol{q})d\delta_i(t_0)/dt + \mathfrak{D}(t_0, \boldsymbol{q})\delta_i(t_0)}{\delta_i(t_0)d\delta_d(t_0)/dt - d\delta_i(t_0)/dt\delta_d(t_0)} . \end{aligned} \tag{10.3.4}$$

For example, if at $t = t_0$ the matter is at rest, i.e., if only density perturbations are specified, then $v(\boldsymbol{q}, t_0) = 0$ and $\mathfrak{D}(t_0, \boldsymbol{q}) = 0$, so that

$$\delta(\boldsymbol{q}, t) = \delta(\boldsymbol{q}, t_0)\frac{\delta_i(t)d\delta_d(t_0)/dt - d\delta_i(t_0)/dt\delta_d(t_0)}{\delta_i(t_0)d\delta_d(t_0)/dt - d\delta_i(t_0)/dt\delta_d(t_0)} . \tag{10.3.5}$$

Consequently, in this simplest of cases, the perturbation grows without a change in form (Doroshkevich and Zel'dovich 1963). Inasmuch as $\delta_i(t)$ grows but $\delta_d(t)$ decays in the course of time, then for $t \gg t_0$ we have $\delta_i(t) \gg \delta_i(t_0)$ and $\delta_d(t) \ll \delta_d(t_0)$. Neglecting the decaying solution, we then obtain, for $t \gg t_0$,

$$\delta(t, \boldsymbol{q}) = \delta(t_0, q)\frac{\delta_i(t)}{\delta_i(t_0)}\left[1 - \left(\frac{d\ln\delta_i/dt}{d\ln\delta_d/dt}\right)_{t_0}\right]^{-1} \tag{10.3.6}$$

if the matter is at rest at the time t_0.

In the general case the initial state is specified by five independent functions of the coordinates—the amplitudes of the density and entropy perturbations and the three components of the velocity vector. However, there is only one large-scale, growing solution. It is also true in the general case that if the initial amplitude of the growing solution is not equal to zero for some special reason, then, in the course of time, the growing part of the

solution turns out to be greater than the remaining parts. Thereupon, one need only be concerned with this one growing solution. In its dependence on time, it grows in proportion to δ_i. The spatial distribution remains similar as time passes. Through formulae (10.3.3) and (10.3.4), it is expressed in terms of the initial (for $t = t_0$) velocity and density distributions.

Note, for example, that if at the initial time $\delta = 0$ and the velocity is purely rotational, $\nabla \cdot v = 0$, then the amplitude of the growing solution equals zero in the linear approximation. Such are the assumptions of the vortex theory of galaxy formation propounded by Ozernoy and Chernin (1967, 1968). Density perturbations then depend on higher powers of the initial velocity. To adequately describe the observed Universe, large velocities are required at the time of recombination.

10.4. A View of Perturbations as Variations in the Parameters of a Solution; the Evolution of Perturbations in a Nonflat ($\Omega_0 \neq 1$) Model

The results obtained in the previous section can provide a new view of perturbations, which also enables us to obtain more easily the solutions in more complicated situations. Thus, let us take the initial perturbations to be of special form. Let $\delta = \text{const.}$ and $\mathfrak{D} = \text{const.}$ at t_0. In other words, let them not depend on coordinates within a large sphere of radius R.

Outside the sphere we set the perturbations equal to zero. It is evident that within the sphere a valid solution is the one describing the evolution of a homogeneous, isotropic model but with different parameters as compared with the outside region. The outer region does not affect the inner region gravitationally by virtue of the spherical symmetry of the problem. The pressure also does not influence the dynamics of the expansion because the perturbations are prescribed on a large scale ($R \gg bt$, where b is the sound velocity).

Consequently, if the unperturbed motion is described by the function $\rho = \rho(t)$, then for a description of the perturbation, we must find a nearby solution $\rho^{(1)}$ of the cosmological equations. Now the function $\delta(t)$ is here defined as $\delta(t) = [\rho^{(1)}(t) - \rho(t)]/\rho(t)$. Since the cosmological equations for $\rho(t)$ contain t only through the derivative, one of the methods for obtaining a solution close to the unperturbed solution consists of carrying out a time translation:

$$\rho^{(1)}(t) = \rho(t + \tau) ; \quad \delta(t) = \tau \frac{d \ln \rho}{dt} = -3\tau H(t) . \qquad (10.4.1)$$

A constant factor is not essential for a solution of the linear equation. Hence one of the solutions (decreasing as the expansion proceeds, and denoted by δ_d in § 10.3) is

$$\delta_d(t) = H(t) . \qquad (10.4.2)$$

One can verify the validity of this solution directly. We know that $dH/dt + H^2 = -4\pi G\rho/3$. Differentiating this equation and using the relation $d\rho/dt = -3H\rho$, we find $d^2H/dt^2 + 2HdH/dt = 4\pi GH\rho$. Therefore, $\delta = H$ satisfies the equation $d^2\delta/dt^2 + 2Hd\delta/dt = 4\pi G\delta\rho$. This last equation coincides with equation (10.3.2) for δ.

In order to find the growing solution $[\delta_i(t)]$, we compare the unperturbed solution $\rho(t)$ with a solution whose density differs:

$$\rho^{(1)}(t) = \rho(t, \Omega^{(1)}) = \rho(t, \Omega + \Omega^*) . \tag{10.4.3}$$

We take the solution in such a way that the times of the singularity [when $\rho = \infty$], for the perturbed solution $\rho^{(1)}(t, \Omega^{(1)})$ and for the unperturbed solution $\rho(t, \Omega)$, coincide. Consequently, both solutions asymptotically coincide at $t \to 0$:

$$\rho^{(1)}(t, \Omega^{(1)}) \approx \rho(t, \Omega) \approx (6\pi Gt^2)^{-1}\,|_{\,t\to 0\ \delta_1\to 0} .$$

The difference between the solutions increases with the passage of time. But the method continues to be applicable even for finite perturbations, when the difference between the solutions is not small. The clearest example is the case in which $\Omega = 1$ for the unperturbed solution; with $\Omega^* > 0$ this gives $\Omega + \Omega^* > 1$ for the perturbed solution. In this case, the expansion of the perturbed region changes to contraction at some instant. Thus, when $t = t_0$, $\rho^{(1)}(t_0) = \infty$, while $\rho(t_0)$ is finite. When $\Omega^* < 0$, the density in the perturbed region varies as $\rho^{(1)} \sim t^{-3}$ for $t \to \infty$. This replaces the law $\rho = (6\pi Gt^2)^{-1}$, which applies to the unperturbed region. Thus, for $\Omega = 1$, perturbations of either sign qualitatively change the result in the limit $t \to \infty$. In linear theory, we can take account of the term first-order with respect to Ω^*, so the result has the form

$$\delta_i(t) = \frac{\partial}{\partial\Omega} \ln \rho(t, \Omega) \tag{10.4.4}$$

for arbitrary $\rho(t, \Omega)$. Expression (10.4.4) gives a growing solution for the perturbation equation.

An examination of the spherically perturbed region is also useful for drawing conclusions from equation (10.3.2), which describes the development of perturbations in another way. Thus, one can write for the spherically symmetric case (with R the radius of the unperturbed region and Δ the change in the radius due to the perturbation)

$$\frac{d^2R}{dt^2} = -\frac{GM}{R^2}; \quad \frac{d^2(R+\Delta)}{dt^2} = -\frac{GM}{(R+\Delta)^2}; \quad \rho = \frac{3}{4\pi}\frac{M}{R^3};$$

$$\frac{d^2\Delta}{dt^2} = \frac{2GM\Delta}{R^3} = \frac{8\pi G}{3}\rho(t)\Delta; \quad \rho^{(1)} = \frac{3}{4\pi}\frac{M}{(R+\Delta)^3}; \tag{10.4.5}$$

$$\delta = -3\frac{\Delta}{R}; \quad R \sim \rho^{-1/3} .$$

Finally,

$$\frac{d^2\delta\rho^{1/3}}{dt^2} = \frac{8\pi G}{3}\delta\rho^{1+2/3}. \tag{10.4.6}$$

This can be transformed to the form (10.3.2) with the help of the equations describing the unperturbed motion. Indeed, this method is more obvious than the long derivation based on Fourier components that was used in § 10.3.

Let us examine in more detail the velocity perturbation due to a change in the density within the given sphere $r < R$. The surrounding matter does not gravitationally influence the matter inside. Hence, each perturbed shell still undergoes Hubble expansion—but governed by a perturbed function $H(t)$. However, the density perturbation within the perturbed region $r < R$ does gravitationally affect the outer region $r > R$. In this latter region the velocity turns out to be perturbed in a way that a simple Hubble law cannot describe: $u_r = H^{(0)} r + A(t)/r^2$. It would not be entirely accurate to maintain that the solution as a whole can be built up out of the solution in the unperturbed region and that in the unperturbed exterior. Nevertheless, one should notice that the perturbed velocity in the outer region is such ($\delta_u \sim r^{-2}$) that its divergence is identically zero. This velocity hence does not lead to a density change. Therefore, if one considers only ρ and $\mathfrak{D} = \nabla \cdot \boldsymbol{u}$, then the external region is unperturbed. When, conversely, we find the velocity field (subject to the condition $\nabla \times \boldsymbol{u} = 0$; i.e., when we find the potential or irrotational part of the velocity from the magnitude of the divergence), then nonzero integrals of the form

$$\boldsymbol{u} = -\nabla \int \frac{\nabla \cdot \boldsymbol{u}(r')}{|r - r'|}\, d^3r'$$

enter the calculation. These integrals give $\boldsymbol{u} \neq 0$ in regions where $\nabla \cdot \boldsymbol{u} \equiv 0$, i.e., in regions outside those with perturbed density.

Inasmuch as the linear theory of the development of the perturbations in the Friedmann model with the equation of state $P = 0$ is so simple and can be constructed immediately, the reader might be wondering why the detailed analysis found above for the self-similar solution was carried out. Is this not like killing an ant with a sledgehammer (or, in Russian, like shooting at sparrows with cannons)? Perhaps, but this examination is useful for the broader goals of understanding the nonlinear situation and also of untangling more complicated cases—including perturbation theory in GTR. Thus the real aim of this section is to provide working tools and ideas about general methods, and not simply to present results and formulae.

10.5 Formulae Describing the Evolution of Perturbations

Above we introduced the quantities, δ, $\mathfrak{D}$, and $\boldsymbol{v}$ and the equations which determine the changes in these quantities. We also discussed considerations

relevant to the solution of these equations. In this section we examine the formulae describing the solutions to these equations, paying particular attention to the dependence on time t or on redshift z. The arbitrary constants appearing in the solutions are chosen so that $\delta = 1$ when $t = t_0$ or $z = 0$. The formulae presented therefore allow one to find out what the perturbations must be at an arbitrary value of z in order that they give $\delta = 1$ today.

The scale of the perturbations l_0 is presented as a wavelength, and also refers to the present time. Thus, we take $k_0 = 2\pi/l_0$. The Hubble constant H_0 and the dimensionless density Ω_0, also evaluated at the present time, are also used in the formulae.

In the simplest case, we have $\Omega_0 = 1$, $H_0 t = \frac{2}{3}(1+z)^{-3/2}$, and $H_0 l_0 = \frac{2}{3}$. Thus,

$$\delta_i = (t/t_0)^{2/3}, \quad \mathfrak{D}_i = -\frac{2}{3t_0}(t/t_0)^{-1/3}, \quad v_i = -\frac{1}{3\pi}\frac{l_0}{t_0}(t/t_0)^{1/3};$$

$$\delta_d = (t/t_0)^{-1}, \quad \mathfrak{D}_d = \frac{1}{t_0}(t/t_0)^{-2}, \quad v_d = \frac{l_0}{2\pi t_0}(t/t_0)^{4/3}; \tag{10.5.1}$$

$$v_{\mathrm{vor}} = v_{\mathrm{vor}\,0}(t/t_0)^{-2/3}.$$

Here v_{vor} is the velocity for a vortex perturbation. If instead of the time one uses the redshift, then

$$\delta_i = (1+z)^{-1}, \quad \mathfrak{D}_i = -H_0(1+z)^{1/2},$$

$$v_i = -\frac{l_0 H_0}{2\pi}(1+z)^{-1/2}; \quad \delta_d = (1+z)^{3/2}, \tag{10.5.2}$$

$$\mathfrak{D}_d = \tfrac{3}{2}H_0(1+z)^3, \quad v_d = \frac{3}{4\pi}H_0 l_0(1+z)^2.$$

Finally, we have $\delta = 0$ and $\mathfrak{D} = 0$ for the vortex perturbations, while the corresponding velocity is

$$v_{\mathrm{vor}} = v_{\mathrm{vor},0}(1+z).$$

Now let us turn to the case $\Omega_0 \neq 1$.

Let us first recall some formulae relating to the unperturbed motion. They are

$$a(t) = a(t_0)(1+z)^{-1};$$

$$H(t) = -(1+z)^{-1}\frac{dz}{dt} = H_0(1+z)(1+\Omega_0 z)^{1/2}; \tag{10.5.3}$$

$$H_0 dt = -(1+z)^{-2}(1+\Omega_0 z)^{-1/2}dz.$$

The normalized, decaying perturbation mode can then be expressed in the form

$$\delta_d = (1+z)(1+\Omega_0 z)^{1/2}, \tag{10.5.4}$$

in correspondence with the results mentioned above. The corresponding velocity perturbations and the divergence of the velocity are

$$v_d = \frac{H_0 l_0}{4\pi}(1+z)(2+\Omega_0+3\Omega_0 z)\,,$$
$$\mathfrak{D}_d = \frac{H_0}{2}(1+z)^2(2+\Omega_0+3\Omega_0 z)\,. \tag{10.5.5}$$

Remember that the functions which grow with increasing z describe quantities that decrease in the course of time, since z itself decreases with increasing time.

It is easy to obtain the growing mode perturbations $\delta_i(t)$ or $\delta_i(z)$ from equation (10.2.6). One easily finds from this equation that for any two solutions $(d\delta_i/dt)\delta_d - \delta_i(d\delta_d/dt) = \text{const.}/a^2$. Thus, if one solution is known (δ_d, say), then for the other we can pass to a first-order equation. Integrating this equation, we find

$$\delta'_i = \delta_d \int \frac{dt}{a^2\delta_d^2} = \delta_d \int a^{-2}\delta_d^{-2}\left(\frac{dt}{dz}\right)dz$$
$$= (1+z)(1+\Omega_0 z)^{1/2}\int_z^\infty (1+z')^{-2}(1+\Omega z')^{-3/2}dz'\,. \tag{10.5.6}$$

This integral is expressible in terms of elementary functions as

$$\delta'_i = \frac{1+2\Omega_0+3\Omega_0 z}{(1-\Omega_0)^2} - \frac{3}{2}\frac{\Omega_0}{(1-\Omega_0)^{5/2}}(1+z)(1+\Omega_0 z)^{1/2}$$
$$\times \ln\left[\frac{(1+\Omega_0 z)^{1/2}+(1-\Omega_0)^{1/2}}{(1+\Omega_0 z)^{1/2}-(1-\Omega_0)^{1/2}}\right]. \tag{10.5.7}$$

This expression is not normalized with respect to the condition that $\delta = 1$ for $z = 0$. To correct this, it is necessary to multiply the expression by the quantity

$$[\delta_i(\Omega, z=0)]^{-1} \approx 1 - \tfrac{3}{2}\Omega_0 \ln\Omega_0 + \tfrac{3}{7}\Omega_0 + \tfrac{15}{14}\Omega_0^2 + \ldots\,. \tag{10.5.8}$$

This approximate formula gives the correct asymptotic form when $\Omega_0 = 1$ and when $\Omega_0 = 0$. Thus, the product of solution (10.5.6) or (10.5.7) and the factor (10.5.8) gives the normalized perturbation $\delta_i(z)$.

One can obtain the asymptotic form for $\Omega_0 \ll 1$ more easily by appealing directly to the integral (10.5.6). Then, for the normalized value of $\delta_i(\Omega_0, z)$, we find these two asymptotic forms: $\delta_i \approx 2/(5\Omega_0 z)$ for $z \gg \Omega_0^{-1}$, and $\delta_i \approx 1 = \text{const.}$ for $z \ll \Omega_0^{-1}$. One can say roughly that, according to the first asymptotic form, δ grows from $z \gg 1$ up to $z_1 \approx 2/(5\Omega_0)$. For $z_1 \geq z \geq 0$, δ_i is constant and equal to unity. Correspondingly, the velocity perturbation occurring with the density perturbation at first grows in proportion to $(1+z)^{-1/2}$ and then falls in proportion to $(1+z)$. We remember

that the term "grow" and "fall" refer to behavior with respect to increasing time t, or decreasing redshift z.

In order to understand such behavior, recall the properties of the unperturbed model. With $\Omega_0 \ll 1$ (more precisely, for $z < z_1$, including $z = 0$), the expansion occurs with an almost constant recession velocity between all pairs of particles. Moreover, gravitation hardly influences the unperturbed motion (see the description of the Milne model in § 2.4). It is therefore natural that gravitation does not influence the development of perturbations during this period. Hence there is no growth of the perturbations, so $\delta = \text{const}$. It is also natural that the velocity falls in inverse proportion to the scale factor.

Proceeding to a description of the past, we can write

$$\rho = \rho_0(1 + z)^3 , \qquad H(t) = u_{12}(t)/r_{12}(t) ,$$

where $u_{12}(t) = u_{12}(t_0)$ and $r_{12}(t) = r_{12}(t_0)/(1 + z)$. Consequently, $H(t) = H_0(1 + z)$.

Let us give a quantitative expression based on these general considerations. Recall, first of all, the regularities applicable to the homogeneous Universe with small Ω_0. Note that the present values are denoted by the subscript zero; these are in distinction to the variable quantities (z), $H(z)$, and $a(z)$, referring to the time in the past characterized by the redshift z. Thus,

$$\begin{aligned} &a(z) = a_0(1 + z) ; \qquad H(z) = H_0(1 + z)(1 + \Omega_0 z)^{1/2} ; \\ &\Omega(z) = \Omega_0(1 + z)(1 + \Omega_0 z)^{-1} . \end{aligned} \tag{10.5.9}$$

The cosmological model that today is almost empty (almost of the Milne form) was almost flat (almost critical) in the past. For example, for $\Omega_0 = 0.2$, then $\Omega(z = 10) = 0.73$ and $\Omega(z = 30) = 0.87$. The law describing perturbation growth, equation (10.5.6) or (10.5.7), correspondingly deviates from the simple laws for the flat model,

$$\delta \sim t^{2/3} \sim (1 + z)^{-1} , \qquad u \sim t^{1/3} \sim (1 + z)^{-1/2} ,$$

only in the approximation applicable to the present.

Formulae (10.5.6) and (10.5.7) and the formulae describing the corresponding peculiar velocity are inconvenient to work with. However, one can propose a simple interpolating formula having satisfactory accuracy over the whole interval of z and asymptotically exact for large z and in the limit when $z \to 0$ and $\Omega_0 \to 0$. This formula is

$$\begin{aligned} \delta_i &= (1 + \tfrac{3}{2}\Omega_0 + \tfrac{5}{2}\Omega_0 z)^{-1}(1 + \tfrac{3}{2}\Omega_0) \\ &= (1 + \tfrac{3}{2}\Omega_0)(1 - \Omega)[(1 - \Omega_0)(1 + \tfrac{3}{2}\Omega)]^{-1} . \end{aligned} \tag{10.5.10}$$

The formula for the peculiar velocity, characteristic of the growing mode, is

$$u_i = \frac{5}{2}\frac{\Omega_0(1+\frac{3}{2}\Omega_0)(1+z)(1+\Omega_0 z)^{1/2}}{(1+\frac{3}{2}\Omega_0+\frac{5}{2}\Omega_0 z)^2}. \qquad (10.5.11)$$

The peculiar velocity of the perturbation achieves its maximum (in the exact calculation using [10.5.7]) when $z = z_{m1} = (0.38 - \Omega_0)/0.62\Omega_0$. One can also pose a question concerning the time when the ratio of the peculiar velocity with respect to the Hubble flow is a maximum. For example, we can choose two particles which are a distance apart equal to half a wavelength. This occurs when $z = z_{m2} = (0.26 - \Omega_0)/0.74\Omega_0$. Both z_{m1} and z_{m2} depend on Ω_0. This is already evident from the observation that, for $\Omega_0 = 1$, the peculiar velocity and the ratio of it to the Hubble velocity grows continuously, so there is no maximum. From general considerations it is clear that for any value of Ω_0, z_{m1} and z_{m2} always correspond (according to the formula for $\Omega(z)$) to the definite, value of instantaneous Ω:

$$\Omega_{m1} = \Omega(z_{m1}) = 0.38\,,$$

$$\Omega_{m2} = \Omega(z_{m2}) = 0.26\,,$$

which are independent of Ω_0, of course.

10.6 Numerical Estimates

We make use of the theory developed above to calculate the evolution of postrecombination ($z \lesssim 1400$) perturbations in the Friedmann model. With the help of the formulae already obtained, one can estimate what the amplitude of the density and velocity perturbations must be so that they can evolve into the present-day galaxies and clusters of galaxies. One can also calculate the characteristic mass corresponding to the Jeans mass. One ought to keep in mind, however, that the linear theory can, in principle, only give answers correct to within an order of magnitude. For example, the mass $M \sim (\lambda/2)^3\rho$ is connected with the perturbation of wavelength λ by dimensional considerations. By virtue of the statistical character of the problem, even if all the perturbations have one wavelength (an unrealistic assumption, of course) the evolution of the perturbations would still lead to a mass spectrum. The calculation of this spectrum is very difficult, though. Moreover, the present parameters of the Universe are known very inaccurately. In particular we do not know the value of the average density. It is therefore necessary to study the range of values $1 \geq \Omega_0 \geq 0.025$.

For definiteness, though, we shall take the set of values $H_0 = 75$ km s^{-1} Mpc^{-1}, $\Omega_0 = 1$, and $\rho_0 = 10^{-29}$ g cm^{-3}. Then, by the well-known formula $\rho = 10^{-29}(1+z)^3$ g cm^{-3}, we find $\rho_{\text{rec}} = 3 \times 10^{-20}$ g cm^{-3} and $T_{\text{rec}} = 3800$ K since $z_{\text{rec}} \approx 1400$ and $t_{\text{rec}} \approx 0.5 \times 10^{13}$ s. At the time of recombina-

tion, the Jeans wavelength and corresponding mass equal

$$\lambda_J = \left(\frac{5\pi}{3}\frac{RT}{\mu G\rho}\right)^{1/2} \approx 8 \times 10^7 \left(\frac{T}{\rho}\right)^{1/2} = 2.8 \times 10^{19}\,\text{cm} \approx 10\,\text{pc}\,,$$
$$M_J \approx \frac{\lambda^3\rho}{8} \approx 3.3 \times 10^{-11} T^{3/2}/\rho^{1/2} M_\odot \approx 5 \times 10^4\, M_\odot\,, \tag{10.6.1}$$

respectively. Here $R = 8.3 \times 10^7$ ergs (K g mole)$^{-1}$ is the gas constant, and μ is the molecular weight ($\mu = 1$ for neutral hydrogen). The length corresponding to $\lambda_{J,\text{rec}}$ changes in accordance with the factor 14 kpc/$(1 + z)$ during the subsequent expansion. Consequently, all objects heavier than globular clusters, which are characterized by masses $M = 10^6\, M_\odot$, are overcritical, i.e., their masses exceed the Jeans mass. Physically, this means that their postrecombination evolution can be considered without taking account of pressure.

In intergalactic space there is no detectable neutral gas. This may indicate that the gas is warmed up and ionized by secondary processes (thereupon $\mu = \frac{1}{2}$) occurring after the formation of galaxies and clusters of galaxies. Let us assume, for example, that the temperature for such ionized gas equals $T = 10^6(1 + z)$; we then obtain

$$\lambda_J = \frac{2.5 \times 10^{25}\,\text{cm}}{(1+z)} = \frac{8}{1+z}\,\text{Mpc}\,, \qquad M_J \approx 10^{13}\, M_\odot\,. \tag{10.6.2}$$

For such a $T(z)$ dependence, the quantity M_J does not depend on z. Under the assumption that $\Omega_0 = 0.025$ and $\rho_0 \approx 2.5 \times 10^{-31}$ g cm^{-3}, the same formulae give

$$\lambda_J = 1.8 \times 10^{20}\,\text{cm}\,, \qquad M_J = 3.1 \times 10^5\, M_\odot \tag{10.6.3}$$

for the neutral gas at recombination. For an ionized gas at a temperature $T < 10^6(1 + z)$ K, the resulting formulae are $\lambda_J < 50(1 + z)^{-1}$ Mpc and $M_J < 6.3 \times 10^{13}\, M_\odot$. Under the assumption that $T \approx 10^6(1 + z)$ K, $\Omega_0 \approx 1$ over large scales (40–60 Mpc). For such scales, the inhomogeneity is already small. But if $\Omega_0 \ll 1$, even this scale is close to that of the Jeans length of the hot ionized gas. We emphasize once again that the expression $T = 10^6(1 + z)$ K is a guess for the late period of small z. During the period of recombination and for the subsequent era, i.e., up to a redshift $z \approx 100$, the heat exchange between the gas and the radiation is large, and high temperatures are improbable. Other considerations show that the temperature is then very low, approximately equal to radiation temperature $2.7(1 + z)$.

Let us now find what the density perturbation δ_{rec} must be at recombination in order to make possible the birth of separate objects at $z = z_b$. (We take z_b to be in the interval from 2 to 10.) At the same time we suppose that

velocity perturbations are absent. We then obtain

$$\delta_b = 1 = A\delta_i(z_b) \; ; \qquad \delta_{rec} = \tfrac{5}{3}A\delta_i(z_{rec}) = \frac{5}{3}\frac{\delta_i(z_{rec})}{\delta_i(z_b)} \; ,$$

with the factor 5/3 arising because $v = 0$ when $z = z_{rec}$. For $\Omega_0 = 1$ and $\delta_i \sim (1 + z)^{-1}$, this gives

$$\delta_{rec} = 1.2 \times 10^{-3}(1 + z_b) \; .$$

In the case $\Omega_0 = 0.025$,

$$\delta_{rec} \approx 2 \times 10^{-2}$$

if $z_b < 20$.

If, at recombination, the density perturbations are small but the velocity perturbations are large, then one can find the magnitude for $\mathfrak{D}_{rec}$ that leads to $\delta(z_b) = 1$. In this case the amplitude of velocity perturbations depends on the scale of the perturbations or on the mass included within the perturbed region.

For $\Omega_0 = 1$, we obtain

$$\mathfrak{D}_{rec} = 2.5 \times 10^{-16}(1 + z_b) \text{ s}^{-1} \; ;$$

$$v_{rec} \approx \mathfrak{D}_{rec}\frac{\lambda_{rec}}{2\pi} \approx \mathfrak{D}_{rec}\frac{\lambda_0}{2\pi z_{rec}} = 3 \times 10^{-20}(1 + z_b)\lambda_0(\text{cm}) \text{ cm s}^{-1}$$

$$= 10^5(1 + z_b)\lambda_0(\text{Mpc}) \text{ cm s}^{-1} = 3 \times 10^{-6}(1 + z_b)\lambda_0(\text{Mpc})c$$

$$= 4 \times 10^5(1 + z_b)(M/10^{12} M_\odot)^{1/3} \text{ cm s}^{-1}$$

$$= 10^{-5}c(1 + z_b)(M/10^{12} M_\odot)^{1/3} \; .$$

If, in correspondence with observations, one takes the largest scale of present-day perturbations as $\sim$70 Mpc and demands that they be created at a redshift $z_b \approx 2$, then $v_{rec} \sim 5.5 \times 10^{-4}\, c$ at recombination. In accordance with this, $v \sim 10^{-2}c$ when $z = z_b = 2$. It is interesting to compare the quantity $\mathfrak{D}$, the divergence of the velocity perturbation, to the divergence of the velocity of the unperturbed motion, $\mathfrak{D}_{rec}^{(0)} = -3H(t) = -2t^{-1}$, at recombination.

For $\Omega_0 = 0.025$, and $z_b < 20$, corresponding calculations give

$$\mathfrak{D}_{rec} = 0.8 \times 10^{-15} \text{ s}^{-1} \; ;$$

$$v_{rec} = 10^{-19}\lambda_0 \text{ s}^{-1} = 0.3 \times 10^6\lambda_0(\text{Mpc}) \text{ cm s}^{-1} = 10^{-5}\lambda_0(\text{Mpc})c$$

$$= 10^6(M/10^{12} M_\odot)^{1/3} \text{ cm s}^{-1} = 0.3 \times 10^{-4}(M/10^{12} M_\odot)^{1/3}c \; .$$

A qualitative difference exists between the case $\Omega_0 = 1$ (or $\Omega_0 > 1$) and the case $\Omega_0 \ll 1$. In the high-density situation ($\Omega_0 \geq 1$), the growth of small perturbations continues until the present time. At some scales the linear

growth of the perturbations might stop at a certain z_b. But this is only true if $\delta = 1$ for this scale when $z = z_b$. Thereafter a nonlinear stage ensues. Observations show, however, that the perturbations are small on large scales (> 60 Mpc). Consequently, on these scales the growth of perturbations continues when $\Omega_0 \geq 1$.

In the low-density case, $\Omega_0 \ll 1$, it is necessary to assume that a much larger part of the matter passes into galaxies and clusters of galaxies by the present time. But the growth of perturbations in low-density models ceases when $z = z_1 \sim 20$ for $\Omega_0 = 0.025$. One is therefore led to assume that the basic mass in galaxies and clusters of galaxies accumulates sufficiently early; in any case for $z \gtrsim 10$. The long wavelength perturbations remain small and do not grow in the period $z < z_1$. To be exact, the density perturbations remain constant, and the velocity perturbations decrease on large scales.

The calculations for $\Omega_0 = 0.025$ outlined above are not exact even under the assumptions that $P = 0$ and that linear theory is applicable. The point is that at recombination the radiation density is $\rho_{\text{rad}} = \epsilon_{\text{rad}}/c^2 = 1.8 \times 10^{-21}$ g cm^{-3} and the matter density is $\rho_m \sim 0.8 \times 10^{-21}$ g cm^{-3}. The radiation does not interact directly with the neutral gas, but the expansion rate of the unperturbed model is larger than it is when the gravitational effects of the radiation are not included. It is also necessary to add the density due to neutrinos and so forth. The resultant gravitational effects are important for some time after recombination. The calculations performed by Monique Guyot and one of the present authors (Guyot and Zel'dovich 1970) show that the effects of taking account of the influence of the radiation density is to require perturbations 2 to 4 times larger than the above estimates.

10.7 The Instability of a Collisionless Self-gravitating Gas

Consider the problem of perturbations of a homogeneous, collisionless gas. Assume for simplicity that the masses of all of the particles are equal and that the velocity distribution is isotropic. We shall carry out the calculation by the Jeans method, i.e., considering perturbations of a time-independent background. In the example of an ordinary gas whose pressure is governed by the equations of hydrodynamics, we know that the incorrectness of such an approach (see § 10.1) does not yield important differences from the result based on the exact theory.

Hence, let us define the bulk density which enters the Poisson equation

$$\nabla^2\phi = 4\pi G\delta\rho \tag{10.7.1}$$

by means of an integral over the phase-space density. We shall denote the particle mass by m, the density of particles in phase space by n, the particle

coordinates by r, and the particle velocity by v. Then

$$\rho = m \int n(r, v) d^3v . \tag{10.7.2}$$

The kinetic equation takes the form

$$\frac{\partial n}{\partial t} + \nabla_r \cdot (nv) + \nabla_{(v)} \cdot (nF) = 0 , \tag{10.7.3}$$

where

$$F = -\nabla\phi .$$

We seek a solution in the form of a plane wave propagating along the x-axis,

$$\Phi = f e^{ikx+\gamma t} , \tag{10.7.4}$$

where Φ represents any of the quantities $\delta\rho$, $\delta n(v)$, ϕ, and F. We denote by means of the same letters the corresponding factors in front of the exponential in formula (10.7.4). We moreover shall indicate unperturbed quantities by the superscript (0). From the kinetic equation we then obtain

$$\begin{aligned} -\delta n(v) &= \frac{4\pi G\delta\rho}{k^2} \frac{ikv_x}{\gamma + ikv_x} \frac{2dn^{(0)}(v^2)}{d(v^2)} \\ &= \frac{8\pi G\delta\rho(k^2v_x^2 + ikv_x\gamma)}{k^2(\gamma^2 + k^2v_x^2)} \frac{dn^{(0)}}{d(v^2)} , \end{aligned} \tag{10.7.5}$$

after canceling the exponential factors. Substituting this expression for δn into the expression for the density,

$$\delta\rho = m \int \delta n d^3v , \tag{10.7.6}$$

we finally obtain an equation for $\gamma(k)$ in integral form:

$$-\frac{8\pi Gm}{k^2} \int \frac{k^2v_x^2 + ikv_x\gamma}{\gamma^2 + k^2v_x^2} \frac{dn^{(0)}}{d(v^2)} d^3v = 1 .$$

In particular, the $k \to 0$ or long-wave limit is

$$\gamma^2 = -8\pi Gm \int v_x^2 \frac{dn^{(0)}}{d(v^2)} d^3v = 4\pi G\rho . \tag{10.7.7}$$

This is the same result that holds for an ordinary gas.

We find the critical wavelength from the condition $\gamma = 0$. We obtain the result

$$k_J^2 = -8\pi Gm \int \frac{dn^{(0)}}{d(v^2)} 4\pi v^2 dv = 4\pi G\rho\langle v^{-2}\rangle . \tag{10.7.8}$$

This formula is also quite similar to the classical Jeans result $k_J^2 = 4\pi G\rho b^{-2}$ with b the speed of sound. In the case of a collisionless gas, a quantity averaged over the distribution in velocity space,

$$\langle v^{-2}\rangle = \frac{\int (n^{(0)}/v^2) d^3v}{\int n^{(0)} d^3v} = \frac{\int n^{(0)} dv}{\int n^{(0)} v^2 dv} , \tag{10.7.9}$$

occurs instead of b^{-2}. For a Maxwellian distribution,

$$b^2 = \frac{5}{3}\frac{kT}{m}, \qquad [3\langle v^{-2}\rangle]^{-1} = \langle v^2\rangle = \frac{3kT}{m}. \tag{10.7.10}$$

The results for an ordinary gas and for a collisionless gas differ somewhat quantitatively in the case of small perturbations. The essential qualitative difference for a collisionless gas—a gas of stars, for example—manifests itself only during the nonlinear stage when an ordinary gas produces a shock wave but stars fly past one another.

The collisionless perturbations of an electron-ion plasma is considered in every plasma physics textbook. The gravitational collisionless instability is treated by Bisnovatyi-Kogan and Zel'dovich (1970), McCone (1971), Magalinskiy and Sil'vestrov (1972), and Sil'vestrov (1974). For the specific case of galactic theory rotating disks, spiral structure, etc., see Mikhailovskiy, Friedman, and Epel'baum (1970), Bisnovatyi-Kogan (1971, 1972*a*), Mikhailovskiy and Friedman (1971, 1973), Bisnovatyi-Kogan and Mikhailovskiy (1973), and Marochnik and Suchkov (1974).

11.1 Conditions for the Growth of Perturbations

In this section, the approach developed in the previous chapter is applied to the study of the early stages during which the matter is completely ionized, the radiation density dominates, and the matter is coupled to the radiation. In the subsequent development, some necessary refinements will be made. Helium recombines earlier than hydrogen. The matter can be considered strongly coupled to the radiation only for wavelengths greater than some definite wavelength that depends on the temperature T. For $\Omega_0 = 1$, there is a short period when the matter is still completely ionized. During this short period, the velocity of sound changes with time according to $b = (c/\sqrt{3})\,[1 + 3\rho_m/4\rho_r]^{-1/2}$. The matter and radiation densities become equal for $z = z_1 \approx 10^4\,\Omega_0$ or $t \approx 2 \times 10^{11}\,\Omega_0^{-2}$ s. Here one ought to write the law for the change of the matter density with time in the form $\rho_m = 0.8 \times 10^6 (t - t_0)^{-2}$ g cm^{-3}, where $t_0 = 2 \times 10^{11}\,\Omega_0^{-2}(1 - 0.135\,\Omega_0^{1/4})$ s. Recombination takes place when $z \approx 1400$; then $T = 3800$ K at $t_r \approx 10^{13}$ s for $\Omega_0 < 0.07$, or $t_r \approx 5 \times 10^{12}\,\Omega_0^{-1/2}$ s for $\Omega_0 > 0.07$. The role of weakly interacting particles will also be examined in more detail.

During this period the Universe is filled with a medium whose equation of state is $P = \epsilon/3$, so the velocity of sound is then $b = c/\sqrt{3}$. In the expanding Universe, $P = \epsilon/3 \sim a^{-4}$, where a is the scale factor. For the unperturbed expansion moreover, $a \sim t^{1/2}$ and $\rho = 3(32\pi G t^2)^{-1} = (4.5 \times 10^5/t^2)$ g cm^{-3}, the latter holding independent of the value of Ω_0. If the ratio of the total energy density of photons, neutrinos, and gravitons to the energy density of photons equals g, then

$$g\sigma T^4 = \rho c^2\,; \quad T = 0.58 \times 10^9 g^{-1/4}\rho^{1/4}\ \mathrm{K} = 1.5 \times 10^{10} g^{-1/4} t^{-1/2}\ \mathrm{K} = 2.7(1 + z)\ \mathrm{K}\,. \tag{11.1.1}$$

The matter density is

$$\rho_m = \Omega_0\rho_c(1 + z)^3 = 2 \times 10^{-29}(\Omega_0 h_0^2)(1 + z)^3\ \mathrm{g\ cm^{-3}} = 3.4(\Omega_0 h_0^2) g^{-3/4} t^{-3/2}\ \mathrm{g\ cm^{-3}}\,. \tag{11.1.2}$$

With the help of the Jeans criterion, let us now find the conditions dividing the regions of stability and instability. We write the perturbations as

$$\delta = \delta_\kappa(t)e^{i\mathbf{K}\cdot\mathbf{q}} = \delta_\kappa(t)e^{i\mathbf{k}\cdot\mathbf{x}},$$

where $\boldsymbol{k} = \kappa a^{-1}(t) = \boldsymbol{k}_0 a^{-1}(t_0)(1+z)$ and $\lambda = 2\pi/k = \lambda_0/(1+z)$. Both k_0 and λ_0 refer to the present time. The Jeans criterion takes the form $b^2k^2 = 4\pi G\rho$. Substituting the values of ρ and b for the radiation-dominated era, we obtain

$$k_J = \frac{3}{2\sqrt{2}}(ct)^{-1}, \qquad \lambda_J = \frac{2\pi}{k_J} = ct\,\frac{4\pi\sqrt{2}}{3}. \tag{11.1.3}$$

The result that the Jeans length λ_J is of order ct (since the sound velocity is of order c) has an obvious meaning. The distance ct is that over which the equalization of the density can occur by virtue of the pressure gradients. This is essential and is physically clear.

The quantities $\lambda_{J,0}$, $k_{J,0}$, and M_J are invariant characteristics of a perturbation, i.e., quantities independent of the expansion. The mass M_J is defined as the amount of matter (without radiation) contained in the volume $(\lambda_J/2)^3$. We must express these quantities in terms of z and t by using the formulae presented above. Doing so, we get

$$\begin{aligned} k_{J\,0} &= k_J(1+z)^{-1} = 6.3\times10^{-21}g^{1/4}t^{-1/2}\ \mathrm{cm}^{-1} \\ &= 1.14\times10^{-30}g^{1/2}(1+z)\ \mathrm{cm}^{-1}; \\ \lambda_{J\,0} &= 10^{21}g^{-1/4}t^{1/2}\ \mathrm{cm} = 5.5\times10^{30}g^{-1/2}(1+z)^{-1}\ \mathrm{cm}; \\ M_J &= \rho_0\left(\frac{\lambda_{J,0}}{2}\right)^3 \approx (\Omega_0h_0^2)g^{-3/4}t^{3/2}\,M_\odot \\ &\approx 10^{29}(\Omega_0h_0^2)g^{-3/2}(1+z)^{-3}\,M_\odot. \end{aligned} \tag{11.1.4}$$

It is convenient to depict the regions of stability and instability on the log t-log M plane of Figures 43 and 44. We take M for the abscissa and t for the ordinate. Formula (11.1.4) for $M_J(t)$ appears as the straight line on these diagrams. It is easy to verify that below and to the right of this line lies the region of instability, while above and to the left lies that of stability.

The limits in the region of very small masses (the lower left part of the curve) interests us little now. It is clear that the theory is valid only before recombination ($t \le t_{\mathrm{rec}}$; the horizontal line in Figs. 43 and 44 represents $t = t_{\mathrm{rec}}$). Figure 43 corresponds to the case when $\Omega_0h_0^2 \approx 0.014$ and $g = 2$. Then the condition $\rho_r > \rho_m$ is fulfilled right up to t_{rec}. Figure 44 corresponds to the value $\Omega_0h_0^2 = 1$. For this case, $\rho_m = \rho_r$ when $z = z_1 \equiv 3\times10^4g^{-1}$. For the general case $\rho_m = \rho_r$ when $z = z_1 \equiv 3\times10^4\,\Omega_0h_0^2g^{-1}$, and during the epoch $z_1 > z > z_{\mathrm{rec}} = 1400$ (if $z_1 > z_{\mathrm{rec}}$), the Jeans mass is constant, $M_J = 3.8\times10^{15}g^{3/2}(\Omega_0h_0^2)^{-2}\,M_\odot$.

After recombination, the considerations of the previous chapter become important. The critical Jeans mass depends only on the pressure of the

matter. This is because the density perturbations of the neutral matter are not accompanied by a change in the radiation density when $t > t_{rec}$. Thus, the very character of the motion connected with the density perturbations differs substantially in the case $t < t_{rec}$ and $t > t_{rec}$. When $t < t_{rec}$, the contraction of the matter is accompanied by the contraction of the radiation. When $t > t_{rec}$, however, the neutral matter contracts against a background of a uniform, unperturbed radiation distribution. The radiation pressure is greater than the matter pressure as measured by the ratio of the number of photons to the number of protons and electrons (before recombination) or to the number of hydrogen atoms (after recombination), a ratio which is about $3 \times 10^7 (\Omega_0 h_0^2)^{-1}$. It is thus not surprising that, when the radiation ceases to influence the matter, the critical Jeans wavelength falls sharply by a factor of $[3 \times 10^7 (\Omega_0 h_0^2)^{-1}]^{1/2}$. Meanwhile, the Jeans mass decreases by the factor $[3 \times 10^7 (\Omega_0 h_0^2)^{-1}]^{3/2}$.

The evolution of a perturbation of a given mass corresponds to the movement along a vertical line in Figures 43 and 44. Let us describe the possibilities.

For $M < M_1$ (region I on the log M scale) perturbations initially in the

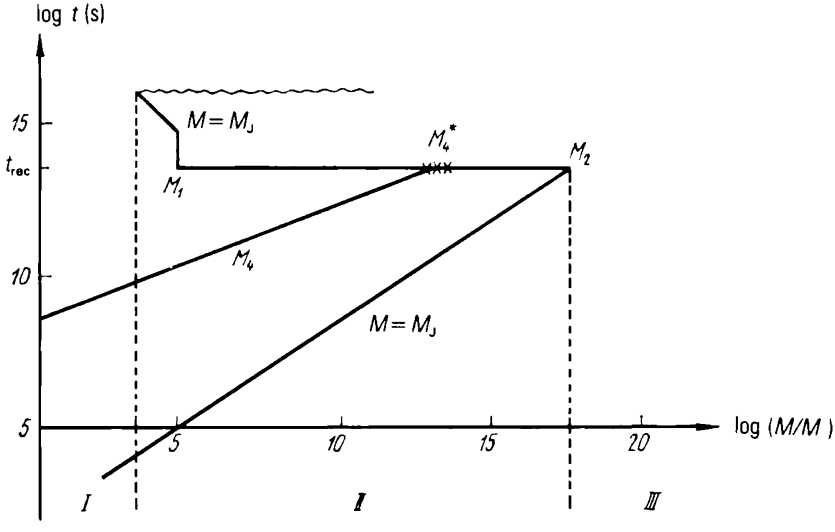


FIG. 43.—The regions of stability and instability on the (log t, log M)-plane for $\Omega = 0.025$. The line $M = M_J$ separates the stability region from the instability region. The line $M = M_4$ is the boundary of the region in which the perturbations effectively dissipate (cf. § 2 of this chapter). Masses $M > M_J$ are located in the instability region, masses $M < M_J$ are located in the stability region, and masses $M < M_4$ exponentially decay with time. The time of recombination $t = t_{rec}$ coincides, for $\Omega = 0.025$, with the time when the matter density and the radiation density are equal. M_1 is the Jeans mass after recombination; it decreases somewhat after $t \approx 10^{15}$ s (cf. § 8 of chap. 14). The wavy line denotes an arbitrarily chosen time when secondary heating of the matter begins (cf. chap. 15). The remaining details are discussed in the text of this section.

region of instability grow, but only up to the time when the vertical line crosses the line corresponding to the mass M_J. Later, the perturbations remain in the region of stability.

For $M_1 < M < M_2$ (region II on the log M scale), the perturbations grow from the time of the singularity right up to the time when the line corresponding to M_J is reached. Thereafter the perturbations fall into the region of stability. We will see later that, in this region, oscillations of constant or decaying amplitude occur. The perturbation continues growing after recombination, when the vertical line crosses the horizontal line corresponding to $t = t_{rec}$.

Finally, for $M > M_2$, the perturbations are always found in the region of instability and therefore grow continuously.

The mass M_2 and corresponding length λ_2 (referred to the present time) are of particular significance. For $\Omega_0 h_0^2 = 0.014$, they are $M_2 = 3 \times 10^{17} M_\odot$ and $\lambda_2 = 930$ Mpc, respectively; for $\Omega_0 h_0^2 = 1$, $M_2 = 3.4 \times 10^{16} M_\odot$ and $\lambda_2 = 125$ Mpc. These values are much greater (particularly for low Ω_0) than those inhomogeneities observable in the Universe today.

What is the rate of growth of perturbations in the instability region, and how do the perturbations behave in the stability region? In answering these questions, we note that one can neglect the role of pressure in the instability region. Therefore, in answer to the first question, we make use of the self-

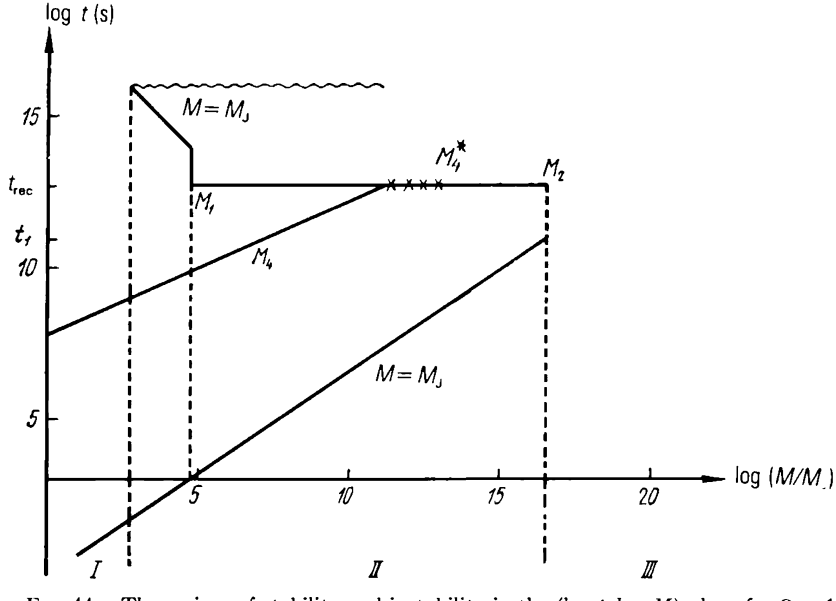


FIG. 44.—The regions of stability and instability in the (log t, log M)-plane for $\Omega = 1$. The notation is the same as in Fig. 43. The matter density and radiation density are equal at the time $t = t_1$; and recombination takes place at the time $t_{rec} > t_1$. During the interval $t_1 \leq t \leq t_{rec}$, the Jeans mass is constant.

similar technique. Let an unperturbed sphere expand in accordance with the equation

$$\frac{1}{2}\left(\frac{dR}{dt}\right)^2 = \frac{B}{R^2}; \quad R = (8B)^{1/4}t^{1/2}. \tag{11.1.5}$$

Let its density be $\rho = D/R^4 = D/(8Bt^2) = 3(32\pi Gt^2)^{-1}$, so that $B = 4\pi GD/3$.

Let the evolution of a perturbed sphere be described by an equation differing from equation (11.1.5) by the small constant term α:

$$\frac{1}{2}\left(\frac{dR'}{dt}\right)^2 = \frac{B}{(R')^2} + \alpha;$$
$$t = \frac{1}{(2B)^{1/2}}\int \frac{R'dR'}{[1 + \alpha(R')^2/B]^{1/2}} \approx \frac{(R')^2}{2(2B)^{1/2}} - \frac{1}{8\sqrt{2}}\frac{\alpha(R')^4}{B^{3/2}}. \tag{11.1.6}$$

In the linear approximation $[\alpha(R')^2/B \ll 1]$, we solve this equation by the method of successive approximations. The first stage gives

$$R' = \left[(8B)^{1/2}t + \frac{\alpha(R')^4}{4B}\right]^{1/2} = (8B)^{1/4}t^{1/2}\left[1 + \frac{\alpha(R')^4}{16(\sqrt{2})B^{3/2}t}\right].$$

We next replace R' in the term on the right by the unperturbed value R, obtaining

$$R' = (8B)^{1/4}t^{1/2}\left[1 + \frac{\alpha t}{(8B)^{1/2}}\right]. \tag{11.1.7}$$

The relative density perturbations are then

$$\delta \equiv \frac{\rho' - \rho}{\rho} = -4\frac{(R' - R)}{R} = -\frac{4\alpha t}{(8B)^{1/2}}$$
$$= \text{const.} \times t = \text{const.}\ (1 + z)^{-2}. \tag{11.1.8}$$

Thus, density perturbations with a scale greater than the Jeans length (i.e., lying in the instability region) grow in proportion to t or $(1 + z)^{-2}$ if the equation of state is $P = \epsilon/3$.

Using the stiffest equation of state, $P = \epsilon$, so $P \sim a^{-6}$ and $z \sim t^{-1/3}$, the same method gives $\delta \sim t^{4/3} \sim z^{-4}$. Note that there is an error in Table XXIII on page 527 of the first Russian edition, *Relyativistskaya Astrofizika* (Relativistic Astrophysics). Decaying density perturbations decrease in time as t^{-1} in all cases. Here, the flat model is taken as the unperturbed solution. In its dependence on the sign of α, the perturbation corresponds to a closed or open cosmological model near the singularity.

The total density ρ and the matter density ρ_m are connected by the nonlinear relation $\rho_m \sim \rho^{3/4}$. For small perturbations, however, the linear relation $\delta_m = (\rho'_m - \rho_m)/\rho_m = \frac{3}{4}\delta$ is satisfied, and δ_m also grows in proportion to t. Remember too that the justification for formulae (11.1.5) and (11.1.6) is presented in the first part of the book and is based upon an exact relativistic treatment of the cosmological model with $P = \epsilon/3$.

In the region of stability the behavior of the perturbations is similar to that of sound waves. In a fluid with $P = \epsilon/3$ and a constant sound velocity, the peculiar velocity and the density of the perturbations are connected by the relation

$$\delta_m = \frac{u}{b} = \tfrac{3}{4}\delta . \tag{11.1.9}$$

The energy of sound waves is a quadratic function of the wave amplitude. After averaging over an interval containing many wavelengths, we write this energy in the form

$$\epsilon_{\text{acous}} = 2\langle\rho\rangle \frac{\langle u\rangle^2}{2} = \tfrac{9}{16}\delta^2\rho b^2 = \frac{3\delta^2}{16}\epsilon . \tag{11.1.10}$$

When the parameters change slowly compared with the oscillation frequency ($\lambda = 2\pi c/\omega < ct$), adiabatic invariants are conserved in an ideal, inviscid fluid. The energy of a wave packet, or the sound energy in a given comoving volume $V \sim z^{-3} \sim t^{3/2}$, is proportional to the frequency, i.e., E/ω is invariant. Moreover, the frequency varies as $b/\lambda \sim z$. Finally,

$$E \equiv \epsilon_{\text{acous}} V \sim \epsilon_{\text{acous}} z^{-3} \sim \delta^2\epsilon z^{-3} \sim \delta^2 z \sim b/\lambda \sim z . \tag{11.1.11}$$

Consequently, the amplitude of the density and velocity perturbations in a sound wave are constant until such time as dissipative processes come into play.

11.2 Dissipative Processes and the Decay of Adiabatic Perturbations

The question of the role of dissipative processes in the evolution of adiabatic perturbations was first examined by Silk (1968). Analogous calculations were done by Doroshkevich (1967), although this work remains unpublished.

Simple dependences and manageable formulae were obtained by Silk for the limiting case when the radiation density is greater than the matter density (RD), when the wavelength is less than the distance to the horizon, $\lambda < ct$, and when the mean free path of the photons, l_γ, is substantially less than ct. The amplitude of the perturbations is considered small; i.e., linear theory is examined. In reality, all of these assumptions are more or less violated near recombination. In several later studies (Peebles and Yu 1970; Chibisov 1972*a*; Field 1973*a*; Weinberg 1971; Bonometto, DeZotti, and Lucchin 1977), the theory of dissipative processes was refined; the numerical results changed noticeably. Nevertheless, it is useful to study first the elementary theory, the region of applicability of which is limited by the above assumptions.

We are thus concerned with a medium whose equation of state is $P = \epsilon/3$, so that the sound velocity is $b = c/\sqrt{3}$. The condition $\lambda < ct$ also means that $\lambda < \lambda_J$, so that one can neglect the role of gravity and consider pure acoustic waves. From a macroscopic point of view, the reason for the

decay of acoustic waves is the viscosity and the thermal conductivity of the medium. In turn, the viscosity and thermal conductivity are related to the diffusion of photons.

The scattering of photons by electrons is governed by the cross section

$$\sigma_T = \frac{8\pi}{3} r_e^2 = \frac{8\pi}{3} \frac{e^4}{m_e^2 c^4} = 0.665 \times 10^{-24} \text{ cm}^2$$

for Thomson scattering. This cross section does not depend on frequency if $h\omega \ll m_e c^2$. Also, scattering by protons can be ignored since m^2 appears in the denominator. For $z < 10^8$, one can also neglect relativistic effects—the creation of e^+e^- pairs and photon-photon scattering. Consequently, in the region of interest to us,

$$l_\gamma = (\sigma_T n_e)^{-1} = 1.25 \times 10^{29} (\Omega_0 h_0^2)^{-1} z^{-3} \text{ (cm)} . \qquad (11.2.1)$$

To compare this with other characteristic lengths, let us find the mass of ordinary matter in a sphere with diameter l_γ. Denoting this mass by M_3 we have

$$M_3 = \tfrac{1}{6}\pi l_\lambda^3 \rho_m = 10^{25} (\Omega_0 h_0^2)^{-2} z^{-6} \, M_\odot . \qquad (11.2.2)$$

At recombination, this mass is

$$M_{3,\text{rec}} \approx 10^7 (\Omega_0 h_0^2)^{-2} \, M_\odot . \qquad (11.2.3)$$

Thus, for masses less than M_3, the decay takes place almost instantaneously during less than one acoustic vibration period. For larger masses $M > M_3$, we may apply the macroscopic concepts of viscosity and thermal conductivity because the photon path length is less than the wavelength of a perturbation. The decay after one oscillation period is not large, but this does not exclude the possibility of a substantial decay under cosmological conditions, since for $\lambda \ll ct$ an acoustic wave undergoes many oscillations over characteristic cosmological times.

The kinematic viscosity, i.e., the ratio $\nu = \eta/e$ of the viscosity to the density, and the thermal diffusivity, i.e., the ratio ϕ of the thermal conductivity to the heat capacity per unit volume, have the same dimensions as the diffusion coefficient. Sometimes the kinematic viscosity is called the coefficient of velocity diffusion, and the thermal diffusivity the coefficient of temperature diffusion. Now under the assumptions made above,

$$\nu = \phi = \tfrac{1}{3} l_\lambda c = 1.25 \times 10^{39} (\Omega_0 h_0^2)^{-1} z^{-3} \text{ cm}^2 \text{ s}^{-1} . \qquad (11.2.4)$$

Notice that if the density of ordinary matter is of order or greater than the radiation density, but the photon density is significantly greater than the density of protons and electrons, then ν decreases by the factor $\rho_r/(\rho_m + \rho_r)$. The ratio ϕ, however, undergoes no change.

We will now form the equation for the oscillation energy of a single monochromatic wave; an energy understood to refer to a unit comoving volume. Suppose that the wave is characterized by the "comoving" wave

vector k_0 that is connected with the evolving (instantaneous) wave vector k by the relation

$$k = k_0(1 + z) . \tag{11.2.5}$$

In the period of interest to us—before recombination—$z > 1400$, so we may neglect unity in comparison to z in this formula. The equation for the energy takes the form

$$\frac{d(E/\omega)}{dt} = -\left(\frac{E}{\omega}\right)\nu k^2 = -\left(\frac{E}{\omega}\right)\frac{A}{z} ,$$
$$A = 10^{39}(\Omega_0 h_0^2)^{-1} k_0^2 \text{ s}^{-1} . \tag{11.2.6}$$

The ratio E/ω appears in the equation because this quantity is conserved for nondissipative expansion, i.e., it is an adiabatic invariant. Dissipation-dependent decay leads to the result

$$\frac{E}{\omega} = \left(\frac{E}{\omega}\right)_0 \exp\left[-A \int_0^t \frac{dt}{z}\right] . \tag{11.2.7}$$

Since $z \sim t^{-1/2}$, the integral converges for $t = 0$. It is therefore convenient to write

$$t = t_0^* z^{-2} , \quad dt = -2t_0^* z^{-3} dz ,$$
$$t_0^* = 3.1 \times 10^{19} g^{-1/2} \text{ s} = 2.2 \times 10^{19} \text{ s} ; \tag{11.2.8}$$
$$\int_0^t \frac{dt}{z} = \tfrac{2}{3} t_0^* z^{-3} .$$

Thus, the decay of a wave is always determined by conditions during the last part of every period considered. This is because the increase in photon mean free path overpowers the increase in the wavelength.

Let us now take for recombination $z = z_{\text{rec}} = 1400$. Note, however, that this is possible only for $\Omega_0 h_0^2 \leq 0.07$, since then $\rho_r \geq \rho_m$ up to z_{rec}. For larger $\Omega_0 h_0^2$ it is necessary to take account of the behavior of the expansion when $z_1 \leq z \leq z_{\text{rec}}$ (z_1 being defined by $\rho_m = \rho_r$) and to take account of the influence of thermal conductivity. With $z = z_{\text{rec}}$, the integral then may be evaluated, and the result takes the form

$$\left(\frac{E}{\omega}\right)_{\text{rec}} = \left(\frac{E}{\omega}\right)_0 \exp\left(-\frac{2t_0^* A}{3z_{\text{rec}}^3}\right) ,$$
$$\delta_{\text{rec}} = \delta_0 \exp\left(-\frac{t_0^* A}{3z_{\text{rec}}^3}\right) . \tag{11.2.9}$$

The quantity in the exponent depends on the wavelength of the perturbation. It is convenient to pass from this dependence to a dependence on the mass corresponding to the given length. Hence

$$\frac{t_0^* A}{3z_{\text{rec}}^3} = \frac{t_0^*}{3z_{\text{rec}}^3} 10^{39}(\Omega_0 h_0^2)^{-1} k_0^2 \approx \frac{k_0^2 10^{49}}{2\Omega_0 h_0^2 g^{1/2}} = \left(\frac{M_4}{M}\right)^{2/3} , \tag{11.2.10a}$$

where

$$M = \rho_m \left(\frac{\lambda}{2}\right)^3 \approx 2.7 \times 10^{61} \Omega_0 h_0^2 k_0^{-3} M_\odot ,$$
$$M_4 \approx \frac{3 \times 10^{12}}{(\Omega_0 h_0^2)^{1/2} g^{3/4}} M_\odot \approx \frac{1.8}{(\Omega_0 h_0^2)^{1/2}} 10^{12} M_\odot , \tag{11.2.10b}$$

for $\Omega_0 h_0^2 \lesssim 0.07$. For $\Omega_0 h_0^2 = 1$ the formulae (11.2.10) are not correct and $M_4 \approx 3.3 \times 10^{11} M_\odot$.

Thus, still another characteristic mass, M_4 appears. For $M > M_4$, one can neglect the decay. For $M < M_4$, however, the decay is substantial and the adiabatic perturbations are damped. In particular, the decay factor is

$$\exp\left[-(0.1)^{-2/3}\right] = 10^{-2} \quad \text{for} \quad M = 0.1 M_4 \tag{11.2.11a}$$

and

$$\exp\left[-(0.01)^{-2/3}\right] = 4.4 \times 10^{-10} \quad \text{for} \quad M = 0.01 M_4 . \tag{11.2.11b}$$

Therefore, the mass M_3 for which the decay takes place after one oscillation does not play a role. Curiously, the wavelength of the perturbation decaying after a cosmological time is, in order of magnitude, the geometric mean of photon mean free path and the characteristic length ct. This is immediately apparent from the condition

$$\nu k^2 \sim \frac{c l_\gamma}{\lambda^2} \sim t^{-1} , \quad \lambda^2 \sim l_\gamma c t . \tag{11.2.12}$$

Correspondingly, $M_4 \sim (M_3 M_J)^{1/2}$, where M_J is the Jeans mass. As was noticed by Silk, the coincidence of the characteristic mass M_4 and the mass of large galaxies is most interesting.

The graph $M_4 = M_4(t)$ is shown in Figures 43 and 44. We shall see (at the end of this section) that M_4 increases very rapidly during the short recombination period and becomes equal to $M_4^* = 10^{13} M_\odot$. It is shown on Figures 43 and 44 by the line with crosses.

One can assume that the mass M_4 is not selected out in the initial perturbation spectrum. Rather, the amplitude of perturbations corresponding to different masses grows or remains constant on the low-mass part of the spectrum. However, dissipative processes eliminate all short-wave adiabatic perturbations; the isolation of objects with masses less than M_4 can occur only by means of the successive fragmentation of larger ($M > M_4$) objects.

It is necessary to keep in mind that, for $\Omega_0 h_0^2 > 0.07$, the condition for radiation domination is violated even before recombination. At the same time, thermal conductivity becomes more important than viscosity. Further, what is most important, the recombination itself does not take place instantaneously. During recombination, the electron density decreases, for example, by a factor of 10^4 upon a change in z by a factor of 1.5. In other words, for a change in ρ_m by a factor of 3, the photon mean free path increases by a factor of 10^4. In agreement with the formulae for the thermal

conductivity and the viscosity, these quantities then grow by the same factor. The decay of acoustic waves is therefore intensified—but intensification of the decay by a factor of 10^4 does not occur.

Beginning with the time when the path length becomes equal to the acoustic wavelength, one cannot make use of the concepts of viscosity and thermal conductivity. For a mass $M = 10^{13}\,M_\odot$, $\Omega_0 h_0^2 = 1$, and $z = z_{\text{rec}}$, this condition is achieved when $n_e = 10^2\ \text{cm}^{-3}$, i.e., when the degree of ionization $x = n_e/n_{\text{H}}$ is of order 0.003. A subsequent decrease in the degree of ionization no longer leads to the intensification of the decay. On the contrary, now the characteristic time for the damping of partially ionized ($x < 1$) matter in a radiation field with energy density ϵ_r is

$$\tau = \frac{3}{4}\frac{cm_p}{\epsilon_r \sigma_{\text{T}} x} = 10^{11}\left(\frac{z_{\text{rec}}}{z}\right)\frac{1}{x}\,\text{s}\,. \tag{11.2.13}$$

For $x < 0.01$, this is longer than the cosmological time. The point leading to the new formulae (11.2.13) instead of (11.2.6) is the following. Practically every interaction of the matter and the radiation is then turned off. Inasmuch as the almost totally neutral ($\sim$99%) matter does not influence the radiation pressure, there hence begins after recombination a period wherein the matter moves subject only to the single force of gravity. Equation (11.2.13) expresses the friction of matter moving on the background of strictly homogeneous and isotropic radiation. For this period, therefore, the initial distribution of the matter density depends on processes occurring during the stage prior to recombination. After a discussion of all of the processes playing a role in the evolution of the Universe, the results of numerical calculations describing this period will be given in the concluding section. The fundamental fact for galaxy formation theory is that there is a characteristic "damping" mass $M^* \approx 10^{13}\,M_\odot$. For $M < M^*$, adiabatic perturbations are damped at the end of the period of recombination (Chibisov 1972*a*, *b;* Peebles and Yu 1970).

11.3 The Interaction of Perturbations with Free Particles

In an examination of the evolution of density perturbations and also of the propagation of gravitational waves, there arises a question concerning the possible role of such particles as neutrinos and high-frequency gravitons. Indeed, we studied the limits on the density ρ_d of such difficult-to-observe particles (DOP) in connection with the theory of the hot model discussed above. Recall the conclusion: $0.5\rho_r < \rho_d < 10\rho_r$. Also examined there was the effect of these particles on the expansion (see § 7.5). Finally, we discussed the indirect influence of DOP on the evolution of perturbations, through the change in the general expansion rate.

In principle, however, there is, in addition, the specific effect of the irreversible gain of energy from the acoustic oscillations that leads to the

heating of DOP. This effect depends on the gravitational interaction of a wave of variable density with DOP. Since DOP constitutes a collisionless gas, the phenomenon considered is similar to Landau damping in Vlasov's theory of the oscillations of a collisionless plasma. We will now give an estimate of the order of magnitude of the effect.

An acoustic wave propagates with a velocity $c/\sqrt{3}$. Meanwhile, the perturbations of the gravitational potential are related to the density perturbation by the equation

$$\nabla^2\phi' = -k^2\phi'_k = 4\pi G\delta\rho = 4\pi G\rho^{(0)}\delta_k\,, \tag{11.3.1}$$

where ϕ' is the perturbed potential, ϕ'_k its Fourier component, and δ_k is the Fourier component of the density perturbation. Among DOP moving at the speed of light in all directions are DOP moving in phase with the wave. For this it is necessary that their velocity be directed at an angle $\cos^{-1}(1/\sqrt{3}) \approx 55°$ with respect to the wave vector k. A fraction of the particles, of order ϕ'/c^2, is entrapped and moves in phase with the wave continuously removing energy. The rate of energy transfer is of order

$$\rho_{\text{capt}} u_{\text{capt}} F\,, \tag{11.3.2a}$$

where

$$\rho_{\text{capt}} = \rho_d \frac{\phi'}{c^2}\,, \qquad u_{\text{capt}} = \frac{c}{\sqrt{3}}\,, \qquad F = -\nabla\phi' = ik\phi'_k\,. \tag{11.3.2b}$$

Thus the rate of energy transfer is of order

$$w = k\rho_d\phi'^2 c^{-1} = (4\pi)^2 G^2 (\rho^{(0)})^2 \delta_k^2 \rho_d c^{-1} k^{-3}\,. \tag{11.3.3}$$

It is necessary to compare this expression with the acoustic energy density

$$\epsilon_{\text{acous}} = \rho^{(0)} u^2 = \rho^{(0)} c^2 \delta^2\,. \tag{11.3.4}$$

We thereby obtain the equation

$$\begin{aligned} \frac{d\epsilon_{\text{acous}}}{dt} &= -\epsilon_{\text{acous}} (4\pi)^2 G^2 \rho^{(0)} \rho_d k^{-3} c^{-3} \\ &= -\epsilon_{\text{acous}} \frac{\rho_d}{\rho^{(0)}} \left(\frac{\lambda_J}{\lambda}\right)^{-4} \omega\,. \end{aligned} \tag{11.3.5}$$

Consequently, the effect is of order unity upon passage through the lower limit of the region of stability t_1. At the moment $\omega \sim t^{-1}$, $\lambda \sim \lambda_J$, and we assume $\rho_d \sim \rho^{(0)}$. Subsequently, however, as Peebles (1973*a*) has shown, the damping is again small. The contribution from low z is negligible since, for $t > t_1$, $\lambda_J/\lambda \sim z^{-1}$, $\omega \sim z$, $t \sim z^{-2}$. Hence, upon crossing the boundary of the region of stability, there occurs a single decrease in the amplitude at that time (or near that time). But the effect of DOP is limited by this transient damping.

On the other hand, before crossing the boundary of the region of stability, when $t \ll t_1$, damping of perturbations evidently does not occur. Of course, it was already noted that one can consider the perturbed region as an independently evolving part of the Universe with different (by comparison

with those of the surrounding matter) initial parameters. At the edges of the perturbed region there are nonzero gradients of the density, the pressure, and the velocity. Moreover, waves travel from the edges deep into the perturbed region—relaxing the gradients and thereby destroying the idyllic picture described above. Now the condition $t < t_1$ must be satisfied if this is not to affect the whole region. But the velocity of these relaxation waves, $c/\sqrt{3}$, differs little from the velocity, c, of DOP. When $t \ll t_1$, therefore, the evolutionary picture of the perturbed region is preserved even in the absence of DOP.

Hence, the general conclusion is that the presence of DOP gives rise to a rather small decrease in the amplitude of oscillations in comparison to that prediction by a theory that does not take account of DOP. This result holds over the whole range of perturbation scales, which are transformed into acoustic oscillations before recombination. The decrease depends hardly at all on the scale and takes place during the period when the perturbation is changing from growing to oscillatory (acoustic).

11.4 Entropy Perturbations

The primordial plasma consists of two different substances: "radiation" and "matter." One might therefore imagine that the ratio of the amount of matter to the amount of radiation can change from point to point. The number of photons per baryon, $\sim T^3/n$, is proportional to the entropy in the radiation-dominated plasma. Therefore, variations of the matter to radiation ratio are called entropy perturbations.

Inhomogeneity of such a type at the early stages of expansion, with $\rho_r \gg \rho_m$, does not give rise to motions of the matter or of the radiation; such inhomogeneities are frozen in. Later, however, after recombination, inhomogeneities in the matter, distributed on a background of homogeneous radiation, lead to motion by the same qualitative mechanism that is operative in the case of adiabatic perturbations.

As a first approximation, let us study density perturbations of the matter specified by

$$\rho_m = \langle \rho_m \rangle [1 + \delta_m(t) e^{i\mathbf{k}\cdot\mathbf{x}}] . \tag{11.4.1}$$

The radiation background is homogeneous, $\rho_r = \text{const}$. We assume that the matter velocity satisfies $u_m \neq 0$, while the background radiation is at rest with respect to the Hubble flow, $u_r = 0$. Now the force of friction, acting per unit mass of the matter, is

$$f = -\frac{4}{3}\frac{\sigma_T}{m_p}\frac{(u_m - u_r)}{c}\rho_r c^2 = -\gamma_f u_m , \qquad \gamma_f = \tfrac{4}{3}\sigma_T \rho_r c/m_p . \tag{11.4.2}$$

The equation for matter conservation and the equation of motion reduce to the form

$$\frac{d\delta_m}{dt} = -iku_m \tag{11.4.3a}$$

and

$$\frac{du_m}{dt} = -ikb_m^2\delta_m - \gamma_f u_m + \frac{i}{k}4\pi G\rho_m\delta_m , \qquad (11.4.3b)$$

respectively. In the equation of motion, the terms on the right-hand side take account of (1) the pressure gradient of the matter itself (but not of the radiation!), proportional to the isothermal sound velocity in the matter ($b_m^2 = RT$), (2) the forces of friction, and (3) matter gravity.

Now in a Jeans-type theory (γ, ρ_m, b_m time independent), the results must assume the form $\delta_m \sim e^{\omega t}$ and $u_m \sim e^{\omega t}$, with ω determined through the equation

$$\omega^2 = 4\pi G\rho_m - k^2 b_m^2 - \gamma_f\omega . \qquad (11.4.4)$$

In the limit of long waves,

$$k \ll k_J^* \equiv \left(\frac{4\pi G\rho_m}{b_m^2}\right)^{1/2}, \quad \lambda \gg \lambda_J^* = \frac{2\pi}{k_J^*} ,$$
$$M \gg M_J \sim \frac{4\times 10^4}{(\Omega_0 h_0^2)^{1/2}} M_\odot , \qquad (11.4.5a)$$

we obtain

$$\omega_1 = \frac{4\pi G\rho_m}{\gamma_f}, \quad \omega_2 = -\gamma_f . \qquad (11.4.5b)$$

It is evident that $\omega_1 > 0$ means instability. But the growth of perturbations takes place very slowly. Substituting numerical values for recombination, we obtain the following:

$$\rho_m = 5\times 10^{-20}\Omega_0 h_0^2 \text{ g cm}^{-3} ; \quad b_m^2 = 5.7\times 10^{11} \text{ cm}^2 \text{ s}^{-2} ;$$
$$\rho_r = 2\times 10^{-21} \text{ g cm}^{-3} ; \quad \sigma_T = 0.665\times 10^{-24} \text{ cm}^2 ;$$
$$\gamma_f = 3\times 10^{-11} \text{ s}^{-1} ; \quad t_{rec} = 1.5\times 10^{13} \text{ s} ; \qquad (11.4.6)$$
$$\omega_1 = 1.4\times 10^{-15}\Omega_0 h_0^2 \text{ s}^{-1} ; \quad \omega_2 = -3\times 10^{-11} \text{ s}^{-1} ;$$
$$\int \omega_1 dt \sim \omega_1 t \sim 2\times 10^{-2}(\Omega_0 h_0^2) ; \quad \omega_2 t = -450 .$$

Although an intensification of the "growing mode" occurs in principle, in practice it is almost totally suppressed because of friction, so that $\exp(\int \omega_1 dt) \leq 1.004$. The second root ω_2 corresponds to the damping of the initial velocity perturbations of the matter which is moving relative to the background of density-perturbation-free radiation. The matter motion decays quickly; since $\omega_2 \sim t^{-2}$, the integral $\int \omega_2 dt$ diverges at small t.

The second limiting case corresponds to short waves $\lambda < \lambda_J^*$, so that $M < 5\times 10^4 M_\odot$. Here both roots are negative: $\omega_1 = -k^2 b_m^2/\gamma_f$ and $\omega_2 = -\gamma_f$. Now the condition $|\omega_1 t|_{rec} = 1$ is fulfilled for $k_{rec} = 3\times 10^{-18}(\Omega_0 h_0^2)$ cm^{-1}. This means that for all masses $M > 15(\Omega_0 h_0^2) M_\odot$ the evolution of matter density perturbations (during the entire period from the singularity to recombination) is completely negligible. In other words, the initial distribution of matter density in comoving space is preserved from

the very early moments (from the singularity) right up to the moment of recombination. Small-scale perturbations satisfying $M < 15\Omega_0 h_0^2 M_\odot$ are smoothed out because of the motion of the matter relative to the radiation. One can, of course, study the smoothing out process in detail; it proceeds in an oscillatory (for $k > \gamma_f/b_m$) or a nonoscillatory (for $k < \gamma_f/b_m$) manner. However, the study of this process would yield no useful information. One curious about it would remind us of the poor doctor who, having found out that his patient has died, is forever interested in knowing whether the sick patient hiccupped before dying.

Entropy perturbations satisfying $M > 15\Omega_0 h_0^2 M_\odot$ do not decay by recombination, and consequently remain in practically the same form as they are initially, i.e., near the singularity, $t \to 0$. Just after recombination, however, neutral gas pressure plays a decisive role in their evolution, for the Jeans mass determined by this pressure is $M_1 = 5 \times 10^4 (\Omega_0 h_0^2)^{-1/2} M_\odot$.

Entropy perturbations involving masses $10^5 M_\odot \leq M \leq M_1$ are preserved (are constant) before recombination, but pass into acoustic oscillations of neutral gas and decay rapidly after recombination. Perturbations involving masses $M > M_1$ are preserved before recombination, and thereafter grow! Indeed, hypotheses exist according to which the formation of objects with a mass of order M_1—either supermassive stars (Doroshkevich, Zel'dovich, and Novikov 1967*a*) or globular clusters (Dicke and Peebles 1968)—results from entropy perturbations.

Notice finally that the idealized picture consisting of entropy perturbations on a background of constant radiation pressure is valid only for a wavelength less than ct_{rec} (the horizon), i.e., for a mass less than $\sim 10^{16}$–$10^{17} M_\odot$. This condition is necessary if the pressure is actually to be equalized.

It is easy to convince oneself that in the limiting case of a wave vector $k \to 0$ (perturbed mass $M \to \infty$), entropy perturbations in a flat model lead only to decaying density perturbations! To see this, let us compare two flat models having different entropy and evolving independently of one another. In the limit $t \to \infty$, the radiation density for any finite entropy becomes small in comparison with the matter density. Consequently, $\rho_m \to (6\pi G t^2)^{-1}$ independently of the entropy. Thus, the relative difference in the density, dependent on the difference in the entropy, tends to zero with the passage of time.

11.5 Rotational Perturbations

Rotational perturbations were examined in the previous chapter. Here we will note the peculiarities in the behavior of rotational perturbations that are connected with (*a*) the change in the mass of radiation in the course of the adiabatic expansion and (*b*) photon viscosity.

Recall that in this chapter a Newtonian examination of the linear problem is being conducted. Nevertheless, we will again and again return to a

consideration of rotational perturbations in the subsequent development in the GTR treatment and in the nonlinear theory, in which turbulent motion plays a role.

To begin, then, what is the evolution law for the rotational velocity in the simplest case of an ideal, inviscid fluid? Consider a sphere of radius R with the matter contained within described by a density ρ and a velocity $u \quad \omega \times r$. The maximum value of the velocity, $u_{\max} = |\omega|R$, is achieved at the equator. The angular momentum of the sphere equals $8\pi\rho R^5\omega/15$. If there is no friction, then this angular momentum is conserved during expansion. During the expansion, $R \sim a(t)$, while $\rho \sim a^{-3}(t)$ in the MD ($P = 0$) case and $\rho \sim a^{-4}(t)$ in the RD ($P = \epsilon/3$) case. Therefore,

$$\omega \sim a^{-2}, \quad u_{\max} \sim a^{-1} \quad \text{when} \quad P = 0; \tag{11.5.1}$$

$$\omega \sim a^{-1}, \quad u_{\max} \sim \text{const.} \quad \text{when} \quad P = \epsilon/3. \tag{11.5.2}$$

Thus, during the early stages the vortex (rotational) velocity is conserved upon expansion!

This law can be obtained even more simply, since pressure gradients do not arise (at least in the linear approximation) for rotational motion. Hence, translational motion in a given comoving volume obeys the same law as does the motion of noninteracting particles, namely

$$pa = \text{const.}, \tag{11.5.3}$$

where p is the momentum:

$$p = Mu = \rho V u. \tag{11.5.4}$$

This again gives

$$\rho \sim a^{-3}, \quad V \sim a^3, \quad M = \text{const.}, \quad u \sim a^{-1} \quad \text{when} \quad P = 0, \tag{11.5.5}$$

$$\rho \sim a^{-4}, \quad V \sim a^3, \quad M \sim a^{-1}, \quad u = \text{const.} \quad \text{when} \quad P = \epsilon/3. \tag{11.5.6}$$

Let us compare the result $u = \text{const.}$ for $P = \epsilon/3$ with the result $u = \text{const.}$ for sound oscillations when $P = \epsilon/3$ (see § 11.1). An ultrarelativistic gas satisfying $P = \epsilon/3$ consists of particles (photons) moving at the speed of light c. Upon expansion, the particle energy decreases, but the particle speed does not. This also leads to the results $M \sim a^{-1}$ and $u = \text{const.}$ for $P = \epsilon/3$.

The viscosity of the primordial plasma leads to the decay of rotational perturbations over small scales. But the rate of decay of the rotational motion and of the sound waves can be distinguished only through a dimensionless factor of order unity. Therefore, on these scales, over which viscosity leads to strong damping of the rotational perturbations before

recombination, the perturbation behavior cannot be sharply distinguished from that obtained by Silk (1968) for sound waves.

This is the situation if radiation dominates ($\rho_r > \rho_m$) up to recombination. In the opposite case ($\rho_r < \rho_m$ at recombination), the kinematic viscosity decreases (see above), but the sound waves still decay—now on account of the thermal conductivity. Since vortex perturbations are not accompanied by a change in the density and the temperature, the thermal conductivity does not influence them.

Hence, for large density, $\Omega_0 \sim 1$, only rotational perturbations involving masses $M < 4 \times 10^{11}\, M_\odot$ decay. On the other hand, adiabatic perturbations decay for masses up to $10^{13}\, M_\odot$. Chibisov (1972*a*, *b*) elucidated this circumstance in connection with the vortex theory.

11.6 Matching Perturbations When the Equation of State Changes

The epoch of recombination is witness to transitional phenomena in the growing perturbations. Sakharov (1965) noticed that these phenomena lead to a peculiar periodic dependence of the final perturbation amplitudes on wavelength. He considered the cold model of the Universe, though qualitatively speaking, the same phenomena occur in the hot model. We will examine this effect in the hot model. In the first approximation, we consider recombination as occurring instantaneously and will study a perturbation with wavelengths between the Jeans length for the neutral gas and the Jeans length in the prerecombination plasma. Recall then that the mass M involved satisfies

$$M_1 = 4 \times 10^4\, M_\odot < M \leq M_2 = 3.4 \times 10^{16}\, M_\odot .$$

This mass interval (the numbers are given for $\Omega_0 h_0^2 = 1$) is very wide. Actually, the part of the interval $M_1 < M < M_4$ is not essential. This is because damping causes the wave amplitude to be small in this region (see § 11.2).

In the short period of time around recombination, one can neglect the expansion (i.e., one can return to the Jeans method of examining the perturbations). Meanwhile, the growing and decaying perturbations depend exponentially on time. Finally, we shall denote the frequency $(4\pi G\rho_{\text{rec}})^{1/2}$ by ω, where ρ_{rec} is the density at recombination. Before recombination, then, the solution for the perturbation has the form

$$\begin{aligned} \delta &= A_k \cos \boldsymbol{k}\cdot\boldsymbol{x} \cos (kbt' + \alpha_k) ; \\ u &= bA_k \sin \boldsymbol{k}\cdot\boldsymbol{x} \sin (kbt' + \alpha_k) ; \\ b &= \frac{c}{\sqrt{3}} \left(1 + \frac{3\rho_m}{4\rho_r}\right)^{-1/2} . \end{aligned} \tag{11.6.1}$$

The time t' is reckoned from the moment of recombination, and $t' \ll t_{rec}$ as a consequence of $M < M_2$.

After recombination,

$$\delta = \cos \mathbf{k}\cdot\mathbf{x}(B_k e^{\omega t'} + C_k e^{-\omega t'}) ;$$
$$u = \sin \mathbf{k}\cdot\mathbf{x}\left(-\frac{\omega}{k} B_k e^{\omega t'} + \frac{\omega}{k} C_k e^{-\omega t'}\right). \tag{11.6.2}$$

The coefficients B_k and C_k can be found by equating the two expressions for δ and u at the time $t' \equiv t - t_{rec} = 0$:

$$B_k + C_k = A_k \cos \alpha_k ; \quad \frac{\omega}{k}(C_k - B_k) = bA_k \sin \alpha_k ;$$
$$B_k = \frac{A_k}{2}\left(\cos \alpha_k - \frac{bk}{\omega} \sin \alpha_k\right)$$
$$= \frac{A_k}{2}\left[1 + \left(\frac{bk}{\omega}\right)^2\right]^{1/2} \cos(\alpha_k + \theta) . \tag{11.6.3}$$
$$\theta = \tan^{-1}(bk/\omega) \approx \tfrac{1}{2}\pi \quad \text{for} \quad M \ll M_2 .$$

An exact solution with an account of expansion requires the replacement of $B_k e^{\omega t'}$ and $C_k e^{-\omega t'}$ by

$$B_k\left(\frac{t}{t_{rec}}\right)^{2/3} \quad \text{and} \quad C_k\left(\frac{t_{rec}}{t}\right)^{2/3}, \tag{11.6.4}$$

respectively, in which case $(bk/\omega) \approx (M_2/M)^{1/3}$. The passage of t through t_{rec} is accompanied by a growth in the growing mode (represented by the B_k term) after recombination in comparison with the density perturbation in a sound wave after recombination. This growth is, in order of magnitude, equal to a factor of $(M_2/M)^{1/3}/2$, where $M_2 = 3.4 \times 10^{16} M_\odot$ and M is the mass determined by the scale of the perturbation. This increase is connected with the matching of the velocity u. For a given density perturbation, the bulk velocity of the matter in a sound wave is greater than in growing perturbations.

Notice the periodic dependence of the final amplitude B_k (for $t > t_{rec}$) on the phase α_k. This dependence is determined by the factor $\cos(\alpha_k + \theta)$ in equation (11.6.3). What is its origin? How does one determine the phase α_k? In order to explain this, it is necessary to return to early stages of the expansion, when the perturbations involving a given mass M under consideration are in the region of instability—below the line $M = M_J$ in Figures 43 and 44.

During this period, some perturbations grow (as t) and others decay. Consider, in the spirit of perturbation theory as applied to the expanding Universe, only the growing mode. It is characterized by

$$\delta = \mathfrak{D}_k t \cos \mathbf{k}\cdot\mathbf{x} ; \quad u = -\frac{\mathfrak{D}_k}{k} \sin \mathbf{k}\cdot\mathbf{x} . \tag{11.6.5}$$

This mode must be joined to sound oscillations of the type (11.6.1). The matching time is defined by the condition

$$t_J = t_1(M/M_2)^{2/3}, \tag{11.6.6}$$

and at this time,

$$kb = (4\pi G\rho)^{1/2} \approx (\tfrac{3}{8})^{1/2} t_J^{-1}. \tag{11.6.7}$$

These perturbations are joined to the perturbations

$$\begin{aligned} \delta &= A_k \cos \boldsymbol{k}\cdot\boldsymbol{x} \cos\left(\int_{t_J}^{t} kbdt + \beta_k\right); \\ u &= bA_k \sin \boldsymbol{k}\cdot\boldsymbol{x} \sin\left(\int_{t_J}^{t} kbdt + \beta_k\right). \end{aligned} \tag{11.6.8}$$

From the matching condition, β_k follows:*

$$\tan\beta_k = -(kbt_J)^{-1} = -(\tfrac{3}{8})^{-1/2}, \quad \beta_k = -1.02. \tag{11.6.9}$$

In order to find the phase α_k at recombination, it is necessary to know the increase in the phase in the course of the oscillations, i.e., in the interval from t_J to t_{rec}. For this, let us compute the expansion-induced change in the wavelength and in the sound velocity. This computation leads to

$$\begin{aligned} \alpha_k &= \beta_k + \int_{t_J}^{t_{rec}} bkdt; \quad k = \kappa t^{-1/2}, \quad \kappa = \text{const.}\, M^{-1/3}; \\ b &= \frac{c}{\sqrt{3}}\left(1 + \frac{3\rho_m}{4\rho_r}\right)^{-1/2}. \end{aligned} \tag{11.6.10}$$

Finally, we obtain in terms of the mass M,

$$\alpha_k = -1.02 + 6.6(M_2/M)^{1/3} \approx 2\pi(M_2/M)^{1/3} - \pi/3. \tag{11.6.11}$$

It is important that α_k depends on the wavelength of the perturbation as determined by the mass M. If one specifies the initial perturbations in the form of a smooth function $\mathfrak{D}_k(q)$ [or $\mathfrak{D}_k(M)$], then in the neutral gas the final amplitude of the growing mode takes the form

$$B_k(M) = \text{const.}\, \mathfrak{D}_k(M) M^{-1/3} \sin\left[2\pi(M_2/M)^{1/3} - \pi/6\right]. \tag{11.6.12}$$

The final amplitude is thus modulated by the last factor, which therefore describes Sakharov's oscillations. Roughly speaking, when the initial stage of growth (for $t < t_J$) of the perturbations is interrupted by the oscillations (for $t_J \le t \le t_{rec}$), but the growth continues after recombination (for $t > t_{rec}$), then the final amplitude of the perturbations is large when there is a whole number of half-waves in the interval $t_J \le t \le t_{rec}$.

The final amplitude equals zero if there is a half-integral number of half-waves in that interval. Amplitudes of zero thus correspond to scales

$$\begin{aligned} \frac{M}{M_2} = {} & 0.630, 0.216, 0.098, 0.053, 0.0315, 0.020, \\ & 0.014, 0.0098, 0.0073, 0.0055. \end{aligned} \tag{11.6.13}$$

* Note by Kip S. Thorne: The numbers and formulae in equations (11.6.9)–(11.6.13) were found by Mr. Ian Redmount to be in error in the Russian edition and have been corrected by him for this English edition.

Of course, all of the numbers are given roughly, but it is a very important exact result that the periodic modulation of the amplitude holds all the way to zero amplitude. See Peebles and Yu (1970) for results of a computer calculation.

It is necessary to explain one additional subtlety. The oscillations were previously taken in the form of standing waves:

$$\begin{aligned} \delta &= A_k \cos \boldsymbol{k}\cdot\boldsymbol{x} \cos (kbt + \alpha_k) ; \\ u &= bA_k \sin \boldsymbol{k}\cdot\boldsymbol{x} \sin (kbt + \alpha_k) . \end{aligned} \tag{11.6.14}$$

But standing waves differ in principle from traveling waves, which are written in the form

$$\begin{aligned} \delta &= A_k \cos (\boldsymbol{k}\cdot\boldsymbol{x} - bkt - \alpha_k) ; \\ u &= bA_k \cos (\boldsymbol{k}\cdot\boldsymbol{x} - bkt - \alpha_k) ; \end{aligned} \tag{11.6.15}$$

or, in the notation

$$\delta = A_k \exp \{i[\boldsymbol{k}\cdot\boldsymbol{x} - bkt - \alpha_k]\} ; \quad u = b\delta .$$

Now, for traveling sound waves, the amplitude of the growing mode would have the form

$$B'_k = A_k[1 + (kb/\omega)^2]^{1/2}$$

after recombination: a form without modulation. If the difference between standing and traveling waves is so large, then why do we work with standing waves? The answer is that upon passing from the region $t < t_J$ of perturbation growth to the region $t > t_J$ of sound waves, we obtain standing waves without any other independent assumptions. If we work with the complex form, then we must make use of the reality condition $A_k^* = A_{-k}$.

In connection with the modulation phenomenon, it is necessary to note that we only found the amplitude of a simple, isolated mode (the amplitude of a Fourier component) with a definite invariant wavenumber $\kappa = ak$. For convenience, we introduced in place of κ a variable with the dimensions of mass, connected with κ by the relation $M = \rho_m(\lambda/2)^3 = \pi^3\rho_m k^{-3}$. This mass M is invariant (does not depend on time) since $\rho_m \sim a^{-3}$. One should not think, however, that the mass spectrum of astronomical objects can directly reflect the Sakharov oscillations. Because of the statistical character of the problem, the superposition of perturbations with given $|\boldsymbol{k}|$ and diverse orientations gives a mass spectrum peaked around the value $M \sim \kappa^{-3}$. On the other hand, the probability of the formation of a bound system with definite mass M depends on the amplitude of the harmonics near the value $\kappa = \text{const.}\ M^{-1/3}$. The distribution of astronomical objects with respect to mass will thus reflect the Sakharov oscillations in a very smoothed-out form only. It is possible that they may not be noticed in a study of the mass spectrum. These questions are examined in more detail in the following chapters.

There is no modulation of the entropy perturbations, which give rise to almost no matter motion before recombination. In this situation, we have simply $B_k = A_k/2$. In principle, therefore, if a detailed study of the perturbation spectrum in the region 10^{16}–$10^{18}\ M_\odot$ shows that modulation is present or, on the other hand, proves that modulation is absent, then we will be able to choose between adiabatic and entropy perturbations for the initial state.

Detailed calculations done by Chibisov (1972*a*, *b*) have shown that there is definitely modulation—though not 100% modulation—even in the entropy-perturbation case, particularly for $\Omega_0 \approx 1$. These calculations take account of the period of recombination and of the period when ρ_m and ρ_r are close in value. For adiabatic perturbations, there is in principle a series of wavelengths for which the theory predicts identically vanishing amplitude for the growing perturbations. This is true for any model parameters (Ω_0, the recombination time, and so forth). In other words, the intensity of Sakharov modulation is 100% for adiabatic perturbations.

Unfortunately, there are still no definite data (except for inequalities) concerning the amplitude of perturbations. Given all the difficulties, it is not clear that we will ever successfully investigate the nature of the initial perturbations using the concept of modulation.

12 GRAVITATIONAL INSTABILITY IN THE GENERAL THEORY OF RELATIVITY

12.1 General Principles and Equations

For a more accurate description of the evolution of perturbations, especially those with a wavelength λ greater than or of the order of ct in a fluid with the equation of state $P = \epsilon/3$, it is necessary to use GTR. First, although several conclusions concerning the development of adiabatic perturbations were obtained in the preceding chapter without using GTR, the strict foundation for these results is based on GTR and therefore will be given below. Second, a deeper understanding of the properties of rotational perturbations can be obtained within the realm of GTR. Finally, only in GTR do perturbations connected with gravitational waves arise.

Questions concerning the evolution of perturbations in GTR were analyzed quite thoroughly in the classic study by Lifshitz (1946). Later, a review of the question was given in several contributions by Lifshitz and Khalatnikov (1963*a*, *b*). The problem takes the following form.

For the unperturbed solution one takes the homogeneous, isotropic Friedmann model. The stress-energy tensor of the material contents is taken to be $T_{ik} = (\epsilon + P)u_i u_k - g_{ik}P$. The energy density and pressure are, in addition, related by an equation of state $P = P(\epsilon)$. In the unperturbed solution, the matter is at rest relative to the reference system, so the components of the four-velocity are $u^0 \equiv 1$ and $u^\alpha \equiv 0$. Small "perturbations" to the model are then specified, i.e., the perturbed metric tensor is written in the form $g_{ik}^{(1)} = g_{ik}^{(0)} + h_{ik}$, where $g_{ik}^{(0)}$ represents the unperturbed metric and h_{ik} the perturbations. The perturbed stress-energy tensor meanwhile takes the form $\epsilon^{(1)} = \epsilon^{(0)} + \delta\epsilon$; δP is determined through the equation of state. Finally, the perturbed velocities u^α are assumed to be small, with u^0 fixed by the identity $u^k u_k = 1$. These expressions are then substituted into Einstein's equations, which relate h_{ik}, $\delta\epsilon$, and u^α to one another and determine their time evolution. Solving these equations constitutes a typical Cauchy problem wherein one must specify the initial perturbations; the solution then gives the time evolution of the perturbations.

Before we proceed, however, we must explain several points. The first concerns the specification of the initial perturbations. At a fixed time $t_0 = \text{const.}$, the perturbations h_{ik}, $\delta\epsilon$, and u^α cannot all be given in an

arbitrary form independent of one another. This is because the Einstein equations impose definite relations among the initial values $(h_{ik})_0$, $(\delta\epsilon)_0$, and $(u^\alpha)_0$. These relations correspond to the inevitable relations among the initial values $(g_{ik})_0$ and $(T_{ik})_0$. That is, these initial values are not independent, but obey relations given by the equations of GTR. The physical reason for the existence of such connections is that the distribution of matter (as expressed through the tensor T_{ik}) determines the gravitational field created by this matter. Since giving the g_{ik} also permits one to calculate the gravitational field, the g_{ik} thus cannot all be given independently. Of course, in addition to the gravitational field that depends on the content, i.e., on T_{ik}, there also exists in GTR a free field (gravitational waves). Therefore, g_{ik} is not determined completely by T_{ik}. In classical Newtonian theory, on the other hand, the connections between T_{ik} and g_{ik} correspond to the Poisson equation, an equation which allows one to find the gravitational potential ϕ from the matter distribution (from the matter density ρ). We emphasize that in Newtonian theory the gravitational field is created only by the matter density ρ, while in GTR it depends both on ρ and on P and u^i. Nevertheless, if one does not put ρ and ϕ equal to zero at spatial infinity, some arbitrariness remains in Newtonian theory too: to ϕ one could add a solution of the empty space equation $\nabla^2\phi = 0$.

In the subsequent development, we shall not discuss these connections between the initial value $(g_{ik})_0$ and $(T_{ik})_0$ in detail. The mathematical aspects are elaborated in the book by Petrov (1966), for example. For our problem, the necessary formulae will be presented at the end of this section. We will make specific simplifications later.

The second point requiring explanation concerns the choice of coordinate system and the physical interpretation of the results obtained in relation to this choice. In GTR the choice of a system of coordinates is arbitrary. This arbitrariness introduces serious problems into the interpretation of the results. Consider an example. Suppose that we in no way change the Friedmann cosmological model. However, suppose also that we examine this model not in the terms of the generally accepted comoving system, but in terms of some other system. We select this other system so that it deviates more and more from the comoving system as time passes. In the formal solution, we shall see that h_{ik}, $\delta\epsilon$, and u^α grow, so that the solution "deviates" from the Friedmann solution more and more. But this "deviation" clearly is in no way physical. Rather, the "deviation" is connected with the choice of reference system. Here one is easily saved from such an "unphysical instability" by returning to the Friedmann reference system. But what should one do when the Friedmann solution includes actual physical perturbations? How does one separate physical deviations from deviations similar to the coordinate deviation just considered?

One of the possible methods for discriminating between them is to choose a comoving reference system in the perturbed solution, i.e., a system

in which the matter is at rest, $u^\alpha = 0$. Keep in mind that we are actually interested in the properties of expanding matter. If, in the comoving reference system, *physical* quantities differ little from the corresponding quantities in the Friedmann solution, then the perturbations are small from the physical point of view. The assertion applies to physical quantities, and not to coordinate-dependent quantities, since one can still select the spatial coordinates in the chosen reference system in various ways (for example, using spherical or cylindrical coordinates). There are similarly many ways to measure the time. (The reader can refer to Volume 1 for more detailed discussions of this coordinate freedom.)

Lifshitz (1946) and subsequently many others pursue another path. A synchronous reference system, for which $g_{00} \equiv 1$ and $g_{0\alpha} \equiv 0$, is taken for the study of the perturbations. After obtaining a solution for the perturbations in this reference system, one can analyze the solution by varying the coordinate system. One can thereby discover which part of a perturbation is physical. That is, one can then find which part describes a change in the density distribution in the comoving system, which part describes the change in the character of the deformation of the medium, and which part describes the appearance in the medium of acceleration forces and absolute rotation. The mathematical apparatus used in describing these quantities can be found in Volume 1. Such an analysis thus permits one to discard the "unphysical" modes.

We shall follow the method pioneered by Lifshitz. For another approach see Hawking (1966*a*). First of all, before passing to a deduction of the formulae, we must make an additional remark. In Lifshitz's work (1946), the matter is described by the hydrodynamic stress-energy tensor T_{ik}. In reality, however, the state is determined not only by the mechanical quantities density and pressure. There exist still other quantities, such as the composition, which also characterize the state. For high density and temperature, when all processes are in thermodynamic equilibrium, the composition is characterized by specifying certain conserved quantities, the "charges": baryon charge, lepton charges, and the entropy. They are all conserved in a comoving volume (as long as dissipative processes are neglected, the entropy is conserved). In this chapter we will ignore the creation and evaporation of primordial black holes (PBH, chap. 24) and the creation of particles near the singularity (chap. 24). Besides this, the "hydrodynamic" stress-energy tensor T_{ik} is most general if viscosity and weakly interacting, "free" particles are accounted for in it! Nevertheless, within the confines of the linear theory, there is reason to think that Lifshitz's results concerning the growth of perturbations do not change qualitatively if these refinements are introduced and that the quantitative changes probably do not exceed 30%.

In the nonlinear approximation, on the other hand, one can expect important changes. As another example, in the presence of particles with

a large mean free path, it is in principle possible that the Helmholtz theorem is violated and that vortex motion develops in the originally vortex-free fluid. Thus, in connection with the development of the theory of the hot Universe, therefore, new problems also arise in the theory of gravitational instability in GTR. The significance of the classical results is not diminished by these complications, however, so let us begin with an account of the classical results.

We shall consider perturbations in regions of space which may be large with respect to ct but small with respect to the radius of curvature a during the period studied. As we know (see chap. 2) $a \gg ct$ always holds during the early stages of evolution. These assumed conditions are equivalent to our considering perturbations in a flat model with the critical matter density. The results obtained will be sufficient for our immediate goal—an analysis of the possible paths of galactic evolution. The general case is studied thoroughly in Lifshitz's paper mentioned above.

In a flat model, where $\rho = \rho_c$ and the three-space sections (t = const.) are Euclidean, the perturbations are decomposed into plane waves, just as is done for perturbations in Newtonian theory. Hence, for the homogeneous and isotropic model (the Friedmann model) with a flat three-space section, we write

$$\begin{aligned} ds^2 &= c^2dt^2 - a^2(t)\delta_{\alpha\beta}dx^\alpha dx^\beta \\ &= a^2(\eta)(d\eta^2 - \delta_{\alpha\beta}dx^\alpha dx^\beta) \ . \end{aligned} \tag{12.1.1}$$

Here, the last equality is obtained by redefining the time coordinate: $a(\eta)d\eta = cdt$. In the subsequent development, then, we shall employ η as the time coordinate. Recall that we found the unperturbed solution for this metric in § 2.1. In this unperturbed solution, the matter is at rest, $u^\alpha = 0$, so $u^0 = a^{-1}(\eta)$. In other words, the reference system describing the unperturbed solution is both synchronous and comoving.

The most general small perturbations of the metric are given by the additional terms $h_{lm}\, dx^l dm^m$, where $\eta = x^0$. But, with the help of a coordinate transformation, one can always return to a synchronous metric, i.e., to a metric not containing off-diagonal terms of the form $dx^0 dx^\alpha$ and with $g_{00} = 1$. Perturbation terms thus remain only in the spatial part of the metric, so we are left with

$$\begin{aligned} ds^2 &= c^2dt^2 - a^2(t)\delta_{\alpha\beta}dv^\alpha dx^\beta - h_{\alpha\beta}dx^\alpha dx^\beta \\ &= a^2(\eta)\{d\eta^2 - [\delta_{\alpha\beta} + h_{\alpha\beta}/a^2(\eta)]dx^\alpha dx^\beta\} \ . \end{aligned} \tag{12.1.2}$$

In formulae (12.1.2), the perturbed metric is given in a synchronous reference system, which is not, generally speaking, comoving if one takes account of the terms in $h_{\alpha\beta}$.

The conditions for being synchronous do not, however, determine a coordinate system completely since there exist coordinate transformations

that leave a synchronous coordinate system synchronous. This arbitrariness makes possible the appearance of the unphysical additions to $h_\alpha{}^\beta$, $\delta\epsilon$, and u_α mentioned above.

We will carry out all computations in the linear approximation, in which the $h_{\alpha\beta}$ are first-order quantities. Now, both deviations of u^0 from unity, and also violations of the equalities $T_0{}^0 = \epsilon$ and $T_\alpha{}^\beta = -P\delta_\alpha{}^\beta$ are possible only in second order because they are proportional to the product $u_\alpha u^\beta$. But $T_0{}^\alpha$ components which are first order now arise due to perturbations. Using the unperturbed metric to raise and lower indices, we obtain

$$h_\alpha{}^\beta = h_\beta{}^\alpha = \frac{1}{a^2} h_{\alpha\beta} , \tag{12.1.3}$$

which constitutes a definition of the quantities $h_\alpha{}^\beta$ with mixed indices. These mixed components $h_\alpha{}^\beta$ (including the one with $\alpha = \beta$) are dimensionless. They determine the magnitude of the metric perturbations in the sense that if $h_\alpha{}^\beta \ll 1$, then the perturbation of the metric is small. This is not true for $h_{\alpha\beta}$ and $h^{\alpha\beta}$, which do have dimensions.

With these remarks in mind, we substitute $g_{00} = a^2$, $g_{0i} = 0$, and $-g_{\alpha\beta} = a^2\delta_{\alpha\beta} + h_{\alpha\beta}$ from the metric (12.1.2) and $T_{ik} = T_{ik}^{(0)} + \delta T_{ik}$ into the Einstein equations to get a system of equations for the perturbations. This system can be written in the following form (Lifshitz and Khalatnikov 1963*a*, *b*) using the synchronous reference system:

$$\tfrac{1}{2}(h^{\delta,\gamma}_{\gamma,\delta} - h_{,\gamma}{}^{,\gamma})\left(1 + 3\frac{dP}{d\epsilon}\right) + \ddot{h} + \dot{h}\frac{\dot{a}}{a}\left(2 + 3\frac{dP}{d\epsilon}\right) = 0 ; \tag{12.1.4}$$

$$(h^{\gamma,\beta}_{\alpha,\gamma} + h^{\beta,\gamma}_{\gamma,\alpha} - h_{,\alpha}{}^{,\beta} - h^{\beta,\gamma}_{\alpha,\gamma}) + \ddot{h}_\alpha{}^\beta + 2\frac{\dot{a}}{a}\dot{h}_\alpha{}^\beta = 0 ; \tag{12.1.5}$$

$$\frac{\delta\epsilon}{\epsilon} = \frac{c^2}{16\pi G\epsilon a^2}\left(h^{\beta,\alpha}_{\alpha,\beta} - h_{,\alpha}{}^{,\alpha} + 2\frac{\dot{a}}{a}\dot{h}\right) ; \tag{12.1.6}$$

$$\frac{u^\alpha}{c} = \frac{c^2}{16\pi G}(h^{,\alpha} - h_\beta{}^{\alpha,\beta})\frac{1}{a^2(P+\epsilon)} . \tag{12.1.7}$$

Here we employ the notation $\dot{a} \equiv \partial a/\partial x^0 = \partial a/\partial\eta$ and $h = h_\alpha{}^\alpha$. Indices after a comma denote simple differentiation with respect to the corresponding coordinate. The quantities u^α are coordinate components of the four-velocity. In the approximation used for these equations, $u^0 = a^{-1}$ is sufficient. Equation (12.1.4) is obtained from the Einstein equation for $R_{\alpha\alpha}$, equation (12.1.5) from that for $R_{\alpha\beta}$, equation (12.1.6) from that for R_{00}, and equation (12.1.7) from that for $R_{0\alpha}$.

12.2 Classification of the Perturbations

We seek the solution of equations (12.1.4–7) for plane-wave perturbations in the following way. In accord with equations (12.1.4) and (12.1.5), we

first find $h_{\alpha\beta}$. Then, by means of the elementary computations indicated in equations (12.1.6) and (12.1.7), we obtain $\delta\epsilon$ and u^α.

The perturbations are considered on the background of a spatially homogeneous and isotropic evolving Universe. Consequently, the solutions, which refer to "elementary," "characteristic," or "fundamental" perturbations (from which arbitrary perturbations are constructed by linear superposition), must be invariant with respect to the same transformations with respect to which the unperturbed solution is identically invariant.

One such transformation is a translation of the coordinates; this transformation results from the homogeneity of space. In nonrelativistic theory it is clear that this requirement leads to the dependence of the solution on the coordinates in the form $e^{i\boldsymbol{\kappa}\cdot\boldsymbol{x}}$. The solution thus depends on a certain three-dimensional vector $\boldsymbol{\kappa}$.

Now let us take account of the second-rank tensor aspect of GTR "solutions." Here, this applies to the tensor $h_{\alpha\beta}$, which is defined in a three-dimensional space. Isotropy signifies the existence of a group of rotations in the three-dimensional space under the action of which all properties are invariant. We shall exploit this fact in using (1) coefficients in front of the exponential and (2) the wave vector to construct a second-rank tensor in the three-dimensional space: the perturbation tensor $h_\alpha{}^\beta$.

Following Lifshitz, we introduce first the scalar $Q = e^{i\boldsymbol{\kappa}\cdot\boldsymbol{x}}$, which leads to a tensor $h_{\alpha\beta}$ in the following way. We can initially compose two tensors from Q:

$$\begin{aligned} Q_\alpha{}^\beta &= \tfrac{1}{3}\delta_\alpha{}^\beta Q\,, \\ P_\alpha{}^\beta &= (\tfrac{1}{3}\delta_\alpha{}^\beta - \kappa_\alpha\kappa^\beta/\kappa^2)Q\,. \end{aligned} \tag{12.2.1}$$

Now, multiply each of the tensors by a factor depending on η and then combine them. We obtain the tensor

$$h_\alpha{}^\beta = \lambda(\eta)P_\alpha{}^\beta + \mu(\eta)Q_\alpha{}^\beta\,.$$

Notice further that one can construct both the vector $\boldsymbol{P} = \boldsymbol{\kappa}Q$ and the scalar Q from such a tensor.

We introduce second the vector $\boldsymbol{\Sigma} = \boldsymbol{S}e^{i\boldsymbol{\kappa}\cdot\boldsymbol{x}}$. But, in order that this not be identical to the preceding case, the vector must be such as to not allow the construction of a scalar from it. We therefore require that $\boldsymbol{S}\perp\boldsymbol{\kappa}$, i.e., $\boldsymbol{S}\cdot\boldsymbol{\kappa} = 0$. Consequently, there are, for a given $\boldsymbol{\kappa}$, two degrees of freedom (two independent components) in $\boldsymbol{S}$. With the help of the vectors $\boldsymbol{S}$ and $\boldsymbol{\kappa}$ one can construct the tensor

$$S_\alpha{}^\beta = \frac{1}{\kappa}(\kappa^\beta S_\alpha + \kappa_\alpha S^\beta)\,. \tag{12.2.2}$$

On multiplying $S_\alpha{}^\beta$ by a factor depending on time, we get $h_{\alpha\beta}$.

Third, we introduce the second-rank tensor

$$\gamma_\alpha{}^\beta = G_\alpha{}^\beta e^{i\boldsymbol{\kappa}\cdot\boldsymbol{x}}\,. \tag{12.2.3}$$

We require that we be unable to construct, with the help of this tensor and the vector κ, a vector or a scalar that can be otherwise obtained through the use of the previous constructions. This forces us to demand that

$$G_\alpha{}^\beta \kappa_\beta = 0 ; \quad G_\alpha{}^\beta \kappa_\beta \kappa^\alpha = 0 . \tag{12.2.4}$$

Remember that these quantities are scalars, vectors, and tensors only in the three-dimensional space.

We will show below that the scalar Q describes perturbations of a scalar quantity—the density—and the vector P describes the corresponding motion, i.e., adiabatic perturbations, and, in particular, sound waves. The vector S describes velocity perturbations, to which no scalar corresponds. This means that $\nabla \cdot u = 0$, i.e., that the velocity field is of a vortex nature. However, $\nabla \times u \neq 0$; if this were not true, then we would have $u \equiv 0$ by the condition that u be finite. Finally, the tensor (12.2.3) describes metric perturbations that are orthogonal to κ, perturbations not accompanied by density and velocity perturbations. These tensor perturbations correspond to gravitational waves propagating through the Universe.

We will discuss in detail below each type of perturbation. Lifshitz (1946) also examined perturbations of the closed ($\Omega_0 > 1$) and the open ($\Omega_0 < 1$) Universe. However, for perturbations whose scale is less than the Hubble radius (at present, $cH_0^{-1} \sim 3 \times 10^3 h_0^{-1}$ Mpc), the differences are small. The extreme case of a perturbation of the whole model is treated in the chapter covering anisotropic models. Here, then, we shall continue to consider perturbations of the flat model only.

12.3 Scalar Perturbations

Let us first study perturbations of the type (12.2.1). In the Lifshitz article it was proved that, for the early stages of expansion, when the equation of state is $P = \epsilon/3$, the solution of the equations (12.1.4–7) can be written in the following form if $\kappa\eta \ll 1$:

$$\begin{gathered} h_\alpha{}^\beta = \frac{3C_1}{\eta} P_\alpha{}^\beta + C_2(Q_\alpha{}^\beta + P_\alpha{}^\beta) ; \\ \frac{\delta\epsilon}{\epsilon} = \frac{\kappa^2}{9}(C_1\eta + C_2\eta^2)Q ; \quad \frac{u^\alpha}{c} = -\frac{i\kappa}{12}\left(3C_1 + C_2\frac{\kappa^2}{9}\eta^3\right)P^\alpha . \end{gathered} \tag{12.3.1}$$

If, on the other hand, $\kappa\eta \gg 1$, then

$$\begin{gathered} h_\alpha{}^\beta = \frac{C}{\kappa^2\eta^2}(P_\alpha{}^\beta - 2Q_\alpha{}^\beta)e^{i\kappa\eta/\sqrt{3}} ; \\ \frac{\delta\epsilon}{\epsilon} = -\frac{C}{9}e^{i\kappa\eta/\sqrt{3}}Q ; \quad \frac{u^\alpha}{c} = \frac{C}{12\sqrt{3}}e^{i\kappa\eta/\sqrt{3}}P^\alpha . \end{gathered} \tag{12.3.2}$$

Here, $Q_\alpha{}^\beta$ and $P_\alpha{}^\beta$ are defined by equations (12.2.1), $P^\alpha = \kappa^\alpha Q/\kappa$, and

$$C = (4\sqrt{3})C_2 + (3i\sqrt{3})\kappa C_1 . \tag{12.3.3}$$

In the solutions obtained, modes corresponding to fictitious perturbations have already been excluded—thereby leaving only physical modes. The former were discussed in § 12.1. The technique for separating out the unphysical modes is given by Lifshitz and Khalatnikov (1963*a*, *b*).

It is now necessary to compare these formal results with the intuitive (i.e., not entirely rigorous) results obtained in the previous chapter. The condition that the metric perturbations be finite near the singularity is that $C_1 = 0$. Taking account of this, the main term in $\delta \equiv \delta\epsilon/\epsilon$ is $C_2\kappa^2\eta^2Q/9$ if $\kappa\eta \ll 1$, $\delta \sim \cos(\kappa\eta/\sqrt{3})$; this solution corresponds to acoustic oscillations with a sound velocity $b = c/\sqrt{3}$, and to constant amplitude density perturbations. Thus, the most important results of the intuitive analysis are supported by the exact theory. This conclusion is also valid for the equation of state $P = 0$, a case studied by Bonnor (1957) using nonrelativistic theory. Curiously, the evolution of theoretical thinking occurred in an unnatural order: the exact theory was developed 10 to 20 years earlier than the intuitive approach.

The exact theory works with metric perturbations. In accordance with equations (12.3.1) and (12.3.2), the mode related to the coefficient C_1 entails infinite ($\sim t^{-1/2}$) metric perturbations as $t \to 0$ (see eq. [12.3.1]). Therefore, if one demands that the perturbations be small at η_0, then $C_1 \ll \eta_0$ is required. For the C_2 mode we have an analogous condition: $C_2 \ll 1$. Consequently, in perturbation theory, in which one assumes that the perturbations are small at the start of the expansion (as $t \to 0$), it is necessary to set $C_1 = 0$ and to consider only the C_2 mode, i.e., the mode satisfying $\delta \sim t$. This mode, which describes finite metric perturbations as $t \to 0$ (the h_β^α in general do not depend on time), is related to density perturbations that are finite at finite times, but that tend to zero in proportion to t as $t \to 0$.

Once this result is known, it is not difficult to obtain it from intuitive considerations also. A measure of the metric perturbations is the dimensionless ratio $\delta\phi/c^2$, where ϕ is the Newtonian gravitational potential. In order of magnitude, $\delta\phi \sim G\delta M/l \sim G\delta\rho l^2 \sim G\rho l^2\delta$. Substituting $\rho \sim t^{-2}$, $l \sim a \sim t^{1/2}$, and $\delta \sim t$ for the $P = \epsilon/3$ case, we find that $\delta\phi/c^2 = \text{const}$. It is moreover easy to verify that $\delta\phi = \text{const}$. for large wavelengths (which correspond to the constancy of the metric perturbations) for any equation of state. For example, if $P = 0$, then $\rho \sim t^{-2}$, $\delta \sim t^{2/3}$, and $l \sim a \sim t^{2/3}$, so $\rho l^2\delta \sim \text{const}$. We emphasize once again that these considerations do not have the force of a proof since Newtonian theory is being used beyond the limit of its applicability, i.e., it is being used in the region where $\phi > c^2$. Moreover, the identification of $\delta\phi/c^2$ with the perturbation of the potential is also not always correct if the potential is expressed in terms of a deforming reference system.

The independence of the metric perturbations with respect to time

during early stages of the expansion is in complete accord with the idea that is underscored by the intuitive analysis of long-wave perturbations. Indeed, we actually demonstrated this independence much earlier. In the first and second chapters the independent evolution of two regions with different initial conditions was considered. Within the confines of GTR, two such regions also differ in their metric and curvature. Nevertheless, in the course of their independent evolution, the difference in the dimensionless characteristic of the metrics is preserved. For example, the ratio of the equatorial to the radial length in any part of the closed model (it is less than 2π) is the same over the entire evolution.

Spherically symmetric perturbations with $\delta > 0$, perturbations which can be represented by superpositions of plane waves, can be studied as a part of a closed model enclosed in a flat model. This interpretation is important for a comparison of the exact theory, based on GTR, with the intuitive analysis. For example, velocity perturbations as given by equation (12.3.1) are proportional to $\kappa^3\eta^3 \sim t^{3/2} \sim \lambda^3$, whereas the intuitive analysis gives $u \sim \lambda\delta/t \sim t^{1/2}$. It would seem that there is an obvious contradiction, but this is not so. The point is that these are the velocities with respect to different reference systems. In the intuitive, Newtonian theory, the velocity is determined with respect to an unperturbed "absolute" reference system. In the exact theory (GTR) the "unperturbed" reference system does not enter in the perturbed solution, so the velocity is already given in the perturbed metric. Now, metric perturbations include a change in the local Hubble constant. Consequently, the main part of the "perturbed velocity" in the intuitive theory is already contained within the metric perturbations. If one therefore computes by both approaches not the velocity, which depends on the choice of reference system, but rather the perturbations of the tensor giving the deformation velocities, one indeed obtains the same results.

The difference in the two approaches to the calculation of u disappears at late stages, when the perturbation wavelength becomes small by comparison to the horizon size, $\lambda \ll ct$. Here the metric perturbations tend to zero, and there is complete correspondence between δ and u in the Newtonian and GTR treatments. This is then a description of sound waves. Evidently a description of short-wavelength sound waves within the confines of GTR is a luxury, but the power and beauty of Lifshitz's method is such that the method can permit itself this luxury. We must emphasize, however, that the decrease in the metric perturbations for $\lambda \ll ct$ applies specifically to the case of an equation of state of a RD plasma, whose sound velocity is $b = c/\sqrt{3}$. Here, the smoothing-out of the perturbations under the influence of a pressure gradient is an important phenomenon. In other words, the "sound" horizon is playing an important role. For the case $P = 0$, there is no sound propagation regime and $b = 0$. Density

perturbations continue to grow for $\lambda < ct$, but the metric perturbations remain constant. In short, the geometry and density are in no way smoothed out in a MD model.

In his pioneering work, Lifshitz emphasized the difference between the power-law growth of δ in the theory based on GTR and the exponential growth in the Jeans theory based on Newtonian mechanics. We now know that the problem is not in the passage from Newtonian theory to GTR, but in the consideration of an evolving Universe. Indeed, $t^n = \exp(\int \omega dt)$, so there is actually no difference. The second very important conclusion drawn by Lifshitz has remained firm, however. This is that to explain the observed picture that includes finite perturbations today (galaxies!), it is sufficient for $\delta \to 0$ as $t \to 0$, whereas the metric perturbation h_α^β must remain nonzero as $t \to 0$. According to an estimate made by Novikov (1964*b*), who examined the specific situation in our Universe, $h_\alpha^\beta \approx 10^{-2}$ to 10^{-3} as $t \to 0$. These values correspond to the scale of clusters of galaxies (and greater). We still know essentially nothing about perturbations on much smaller scales. This estimate changed little when account was taken of the hot Universe. The Universe, therefore, which near the singularity asymptotically approaches strict homogeneity with respect to all parameters (density, entropy, and metric), is incompatible with the presently observed picture! Entropy perturbations, however, are not related to metric perturbations. In particular, they are compatible with $h_\alpha^\beta = 0$ when $t = 0$ even if $\delta\rho_m/\rho_m$ remains finite as $t \to 0$.

In conclusion, we present the solution for the first type of perturbation. This solution is valid in the late stages when $p = 0$. Since there is no rotation, for $p = 0$ the synchronous reference system can also be comoving. The evolution of these perturbations is determined by the formulae

$$\text{if} \quad \kappa\eta \ll 1: \quad h_\alpha^\beta = C_1(P_\alpha^\beta + Q_\alpha^\beta) + \frac{2\kappa^2 C_2}{\eta^3}(P_\alpha^\beta - Q_\alpha^\beta)\,; \tag{12.3.4}$$

$$\text{if} \quad 1 \ll \kappa\eta \ll \kappa: \quad h_\alpha^\beta = \left(\frac{C_1}{15}\eta^2 + \frac{2C_2}{\eta^3}\right)\kappa^2(P_\alpha^\beta - Q_\alpha^\beta)\,. \tag{12.3.5}$$

In both limits we have

$$\frac{\delta\rho}{\rho} = \left(\frac{C_1\eta^2}{30} + \frac{C_2}{\eta^3}\right)\kappa^2 Q\,. \tag{12.3.6}$$

If $\Omega_0 < 1$, then, for $\eta \gg 1$, $a \sim t \sim e^\eta$ and

$$h_\alpha^\beta = [2C_1\kappa^2(\tfrac{1}{3} - 2\eta e^{-\eta}) + C_2 e^{-\eta}](P_\alpha^\beta - Q_\alpha^\beta)\,, \tag{12.3.7}$$

$$\frac{\delta\rho}{\rho} = \left(\frac{C_1}{3} + \frac{C_2}{4}e^{-\eta}\right)Q\,. \tag{12.3.8}$$

This last result demonstrates that $\delta\rho/\rho \to \text{const.} \neq 0$ where $t \to \infty$ (for $\Omega_0 < 1$).

12.4 Vector (Rotational) Perturbations

Let us consider the second type of metric perturbation (12.2.2). The evolution of these rotational perturbations is determined by the formulae

$$h_\alpha^\beta = \sigma(\eta)(S_{\alpha k}{}^{\beta} + \kappa_\alpha S^\beta)e^{i\mathbf{k}\cdot\mathbf{x}}, \quad S_{\alpha k}{}^{\alpha} = 0; \tag{12.4.1}$$

$$\frac{u^\alpha}{c} = \frac{-i\kappa\dot{\sigma}}{2a^2(\epsilon + P)} S^\alpha e^{i\mathbf{k}\cdot\mathbf{x}}; \quad \sigma = \sigma_0 \int \frac{d\eta}{a^2}. \tag{12.4.2}$$

For $P = \epsilon/3$, $a \sim \eta$ and $\epsilon \sim \rho \sim \eta^{-4}$, so we have $\sigma = -C_3/\eta$, $\delta = 0$, and $u^\alpha = -(i\kappa C_3/8) S^\alpha e^{i\mathbf{k}\cdot\mathbf{x}}$. The main result of the intuitive analysis—the time independence of the velocity perturbations—is therefore confirmed. This independence holds for $\kappa\eta < 1$ as well as for $\kappa\eta > 1$, i.e., for $\lambda < ct$ and $\lambda > ct$. But the growth of the metric perturbations ($\sigma \to \infty$) for $\eta \to 0$ is unexpected. This leads to the conclusion that initial perturbations of the rotational type are incompatible with a cosmological solution of the Friedmann type (Ozernoy and Chernin 1967, 1968; Zel'dovich and Novikov 1970). The exact conclusion is that rotational perturbations lead to an asymptotic solution near the singularity that differs from a Friedmann solution.

The result is even more unexpected because it seems natural in the $\delta = 0$ case to take the ratio u/c as a measure of the deviation of the perturbations from the Friedmann solution. But here this ratio remains constant as $t \to 0$. This different asymptotic behavior of the dimensionless quantities u/c and σ requires an explanation. Recall that u^α, as determined in accordance with the Lifshitz method (see eqs. [12.4.1] and [12.4.2]), is in a certain sense not the complete velocity. As we have repeatedly emphasized, the notion of velocity is connected with the choice of reference system and is therefore relative. The most important absolute characteristics of perturbations are, instead, the absolute rotation of the matter elements and the change in character of the deformation of the matter. Rotation is absolute relative to a local inertial frame in which there are no Coriolis and centrifugal forces. Such a system is defined locally with the help of gyroscopes that bear no relation to distant galaxies. As is obvious, we do not shrink from non-Machian concepts. But even though the velocity u^α completely describes the rotation of elements of the medium in the case under consideration, this vector does not completely describe the deformation of the medium. This is because of the lack of rotation of the reference system itself. The system employed here is synchronous, i.e., $g_{0\alpha} \equiv 0$, a fact which signifies the absence of rotation of the reference system (see Vol. 1). Though all of the rotation of the medium is described by its rotation within the reference system, the reference system is itself deforming, and it is necessary to take this into account. After calculating the deformation of the medium with respect to the reference system, it is necessary to account for the deformation of the perturbed reference system. The deformation

of the medium is thus a combination of the deformation caused by motion relative to the reference system and the deformation of the reference system itself. Hence, the velocity determined by equation (12.4.2) corresponds to translational motion (for example, motion along the y-axis with an x-dependent velocity) that combines both the rotation and the deformation of a fluid element.

We recall now how the relative motion of neighboring fluid elements is described. A systematic construction of mechanics in GTR and based on the notion of deformation was carried out by Zel'manov (1959*b*). See Volume 1 for a discussion of this point.

First introduce a coordinate system moving together with a selected particle and located so that the origin of the coordinates coincides with the particle. We then demand that the coordinate system be locally Euclidean and locally inertial. If it is assumed that $u\ (x = 0,\ t = 0) = 0$, it then follows that one can write the velocity of neighboring particles at nearby times as

$$u_\alpha(x, t) = \mathcal{F}_\alpha t + \mathcal{D}_{\alpha\beta}{}^{(0)} x^\beta .$$

This expression is correct to first order in the small quantities x and t. The three-dimensional vector $\mathcal{F}_\alpha$ is that part of the fluid acceleration dependent on nongravitational forces (in hydrodynamics, this force equals the gradient of the pressure). The second-rank tensor $\mathcal{D}_{\alpha\beta}{}^{(0)}$, defined for three-dimensional space, decomposes into two parts, an antisymmetric part $A_{\alpha\beta} = \frac{1}{2}(\mathcal{D}_{\alpha\beta}{}^{(0)} - \mathcal{D}_{\beta\alpha}{}^{(0)})$ and a symmetric part $\mathcal{D}_{\alpha\beta} = \frac{1}{2}(\mathcal{D}_{\alpha\beta}{}^{(0)} + \mathcal{D}_{\beta\alpha}{}^{(0)})$. The antisymmetric part describes the fluid rotation in that the angular velocity is given by the axial velocity $\Omega_\gamma = \epsilon_{\gamma\alpha\beta} A^{\alpha\beta}$, with $\epsilon_{\gamma\alpha\beta}$ the completely antisymmetric unit tensor.

Experimentally, the angular velocity can be found in the comoving system by measuring the Coriolis acceleration. The deformation of a fluid element is completely determined by the symmetric part $\mathcal{D}_{\alpha\beta}$. Notice a theoretical difference between $A_{\alpha\beta}$ and $\mathcal{F}_\alpha$ on the one hand and $\mathcal{D}_{\alpha\beta}$ on the other. The measurement of $A_{\alpha\beta}$ and $\mathcal{F}_\alpha$ requires the introduction of free test bodies and of gyroscopes. For the determination of $\mathcal{D}_{\alpha\beta}$, however, it is sufficient to measure the distance between pairs of fiducial particles of the medium. This does not mean that $\mathcal{F}_\alpha$ and $A_{\alpha\beta}$ are any less real. In particular, knowing the equations of mechanics and the fluid equation of state, and measuring the pressure and density distributions and also $\mathcal{D}_{\alpha\beta}$ and $\mathcal{D}_{\alpha\beta}{}^{(0)}$, one can also determine $\mathcal{F}_\alpha$ and $A_{\alpha\beta}$ without the help of hypothetical test bodies.

The trace of the tensor, i.e., $\mathcal{D} \equiv \mathcal{D}_{11} + \mathcal{D}_{22} + \mathcal{D}_{33}$, a quantity independent of the choice of axes, determines the change in the specific volume of the fluid under a deformation: $\dot{V}/V = -\dot{n}/n = \mathcal{D}$, where n is the number density of conserved particles. The remaining, trace-free part of the tensor $\pi_{\alpha\beta} = \mathcal{D}_{\alpha\beta} - \frac{1}{3}\delta_{\alpha\beta}\mathcal{D}$, characterizes the deformation according to which a

sphere is transformed into a triaxial ellipsoid having the same volume. The tensor $\pi_{\alpha\beta}$ has five independent components. The number five corresponds to the two ratios of axes of the ellipsoid that characterize its shape and the three angles characterizing its spatial orientation. On the other hand, one can consider an ellipsoid as the result of a perturbation of a spherical surface by quadrupole functions characterized by spin $l = 2$. The number of such independent functions is $2l + 1 = 5$. Now the formulae presented above require the metric to be Galilean ($g_{00} = 1$; $g_{11} = g_{22} = g_{33} = -1$; $g_{ik} = 0$ for $i \neq k$) at the given instant. The formulae which would permit one to calculate the velocity-of-deformation tensor $\mathfrak{D}_{\alpha\beta}$, as well as the quantities $\mathfrak{F}_\alpha$ and $A_{\alpha\beta}$, are presented in Volume 1. Such formulae are of course given in terms of the quantities $g_{ik} = g_{ik}\ (x^0, x^1, x^2, x^3)$ and dg_{ik}/dt for a general metric. We require here only a particular case of one of these formulae. Before writing it, though, let us make an additional remark about the problem under consideration.

We want to compute the deformation of an element of the medium in the perturbed problem. After performing the computation, we will be convinced that rotational perturbations lead to substantial anisotropic deformation of matter elements—in distinction to the strictly isotropic expansion that occurs in the unperturbed Friedmann Universe. In particular, this anisotropy of the deformation grows rapidly in the past (as $t \to 0$) and is described by the growth of $h_\alpha{}^\beta$ as the singularity is approached. The anisotropy grows in spite of the continued small size of the rotational velocities. At the end of this section, we will present a physical explanation of this situation.

Hence, let us work at a computation of the deformation of an element of the medium. As was already emphasized, the total deformation equals the sum of the deformations of the reference system and the deformation caused by the motion of the medium in this reference system. Now the general formula required for the calculation of $\mathfrak{D}_{\alpha\beta}$ in a synchronous reference system (see eq. [1.6.4] in Vol. 1) is written in the form

$$\begin{aligned} \mathfrak{D}_{\alpha\beta} &= -\frac{1}{2}\frac{c}{g_{00}{}^{1/2}}\frac{\partial g_{\alpha\beta}}{\partial x^0} = -\frac{1}{2}\frac{\partial}{\partial t}\left(g_{\alpha\beta}{}^{(0)} + h_{\alpha\beta}\right) \\ &= \mathfrak{D}_{\alpha\beta}{}^{(0)} - \frac{1}{2}\frac{d}{dt}\left[\sigma(t)a^2(t)\right]S_{\alpha\beta}\,. \end{aligned} \tag{12.4.3}$$

Here $\mathfrak{D}_{\alpha\beta}{}^{(0)}$ is the tensor of deformation velocities in the reference system describing the unperturbed solution. In place of $h_{\alpha\beta}$ is substituted its value for vortex perturbations [the factor $a^2(t)$ appears upon lowering an index in $h_\alpha{}^\beta$]. Meanwhile, $S_{\alpha\beta}$ does not depend on t. We shall now calculate the anisotropic part of the tensor $\pi_\alpha{}^\beta = \mathfrak{D}_\alpha{}^\beta - \frac{1}{3}\mathfrak{D}g_\alpha{}^\beta$. To order of magnitude, the nonzero terms are

$$|\pi_\alpha{}^\beta| \approx \left[-\frac{\sigma}{t} + \frac{1}{t}\frac{d(\sigma t)}{dt}\right] = \frac{d\sigma}{dt}\,, \tag{12.4.4}$$

where we have used the relation $a(t) \sim t^{1/2}$ for $P = \epsilon/3$. The quantities $\pi_\alpha{}^\beta$ are the deviations in the various directions of the Hubble parameter from the average Hubble parameter $H = \frac{1}{3}\mathfrak{D} = (2t)^{-1}$. The relative size of the deviations in the various directions is then

$$\frac{|\pi_\alpha{}^\beta|}{H} \approx \frac{d\sigma}{dt}\, t \approx \frac{C_3}{t^{1/2}}. \tag{12.4.5}$$

This quantity equals $\sigma(t)$. Its growth as $t \to 0$ is a good measure of the growth of the relative size of the deformation anisotropy.

The second part of the deformation, that connected with the motions, can be computed as the derivative of the velocity with respect to the coordinates. As we saw above, however, the relative velocity is small and remains constant. Moreover, the relative magnitude of the deviations from isotropy of the deformation of the medium does not grow in the past. It is therefore clear, even without calculations, that one can neglect this contribution to the deformation of the medium as $t \to 0$ in comparison with that from equation (12.4.5). Hence, as $t \to 0$, the total deformation anisotropy of the medium is described by equation (12.4.3). From this formula it is apparent that, as $t \to 0$, the deviation of a vortex-type perturbation from the Friedmann background is directly connected not with the rotational velocities of the matter, but with the deformation anisotropies caused by these velocities. This effect is purely a relativistic effect.

At first glance it seems that one can avoid this deformation if one takes for the perturbation rigid-body rotation on a given Lagrangian scale. Under this condition, the deformation is totally absent in the Newtonian approximation. It would seem that the rotation velocity v, which remains constant and small, cannot lead to relativistic effects. This is not so, however, because a perturbation of any scale turns out to be located deep inside the gravitational radius corresponding to the mass which the perturbation encompasses. The whole problem therefore becomes substantially relativistic, and even a small velocity leads to strong effects. That is to say, in addition to the small parameter v/c, a still larger parameter l/ct (l is the scale length of the perturbation) enters the problem, and the relativistic effects are determined by the product of these two parameters. Formula (12.4.5) can thus be rewritten as

$$\frac{\pi_\alpha{}^\beta}{H} \approx \left(\frac{v}{c}\right)\left(\frac{l}{ct}\right). \tag{12.4.6}$$

In the problem under consideration, with rigid-body rotation, perturbations are identically absent only along the rotation axis, where the velocity v vanishes.

Arbitrary perturbations of the vortex type remain small as $t \to 0$, i.e., until such time as $\pi_\alpha{}^\beta/H < 1$. It is thus clear that for each wavelength, a corresponding perturbation ensues in the past at a moment when

$|\pi_\alpha^\beta|$ is comparable with H. In our case, the scale length is proportional to $t^{1/2}$. Therefore, in accord with formula (12.4.6), the critical time when $|\pi_\alpha^\beta|/H \approx 1$ satisfies $(v/c)\,[l_0/(c t_{\rm crit}^{1/2} t_0^{1/2})] = 1$, so

$$t_{\rm crit} = \left(\frac{l_0^2}{c^2}\right)\left(\frac{v^2}{c^2 l_0}\right). \tag{12.4.7}$$

During the period $t < t_{\rm crit}$, the expansion differs strongly from Hubble expansion. In any discussion of galaxy formation in the vortex theory, these conclusions must be taken into account.

In essence, the important difference from the unperturbed model is already evident from the tendency of the dimensionless metric perturbations to diverge as $\eta^{-1} \approx t^{-1/2}$, when $t \to 0$. But the analysis of the deformations we have presented helps one to understand in a detailed way, from the point of view of a local observer, the character of the deviations from the unperturbed model. The objective character of the difficulties, which are not removed by any possible coordinate transformation, is thereby proved with certainty.

To conclude the section, we present the formulae describing vortex perturbations at a late stage for the equation of state $P = 0$: For $\eta \ll 1$,

$$\sigma = -\frac{8C_3}{3\eta^3}, \qquad \frac{u^\alpha}{c} = -\frac{i\kappa C_3}{6\eta^2} S^\alpha e^{i\mathbf{\kappa}\cdot\mathbf{x}}. \tag{12.4.8}$$

For $\eta \gg 1$,

$$\sigma = -4C_3 e^{-\eta}, \qquad \frac{u^\alpha}{c} = -\frac{i\kappa C_1}{6} e^{-\eta} S^\alpha e^{i\mathbf{\kappa}\cdot\mathbf{x}}. \tag{12.4.9}$$

In both cases $u^\alpha \sim a^{-1}$.

12.5 Tensor Perturbations and Gravitational Waves

The solution corresponding to a tensor perturbation is determined by the formulae

$$h_\alpha^\beta = \nu(t) G_\alpha^\beta e^{i\mathbf{\kappa}\cdot\mathbf{x}}, \qquad G_\alpha^\beta \kappa^\alpha = 0, \qquad \delta\rho \equiv 0, \qquad u^\alpha \equiv 0. \tag{12.5.1}$$

Here, G_α^β is a constant (independent of the coordinates and the time) tensor first introduced in § 12.2. For a given wave vector κ, there are two linearly independent tensors G. They correspond to the two polarizations of gravitational waves. The equations of GTR for a relativistic gas with $P = \epsilon/3$ give the solution

$$\nu = C a^{-1} e^{i\kappa\eta} = \eta^{-1}[C_4 \cos(\kappa\eta) + C_5 \sin(\kappa\eta)]. \tag{12.5.2}$$

If the wavelength is less than ct, the solution describes a wavelike gravitational field. It is thus clear that in the course of the expansion there occurs a time t when the length becomes less than the size of the horizon. Thereafter, one can compute by means of the usual formulae the energy density of the gravitational waves (see, e.g., Landau and Lifshitz 1962).

Now the amplitude h of a wave is

$$h \approx \frac{C_{4,5}}{\eta} \sim (1 + z) \sim a^{-1} ,$$

where $C_{4,5} \equiv \text{Max}\ \{|C_4|, |C_5|\}$ and

$$\epsilon_\eta = \frac{\pi c^4}{8G}\left(\frac{h}{\lambda}\right)^2 \sim a^{-4} \sim (1 + z)^4 . \qquad (12.5.3)$$

This is the obvious result describing the change in energy density ϵ_η stored in gravitational waves. Moreover, the relation between the time and the coordinate dependence corresponds to that of a wave solution, with the wave velocity equal to c. As was noticed by Lifshitz, velocity and density perturbations are not connected with gravitational waves.

Actually, the velocity of the matter relative to the perturbed system of coordinates is identically equal to zero in the field of a gravitational wave. However, the relative velocities of particles do change in the presence of a gravitational wave. In correspondence with the formulae of the previous section, upon passage of a wave a sphere of unit radius is transformed into an ellipsoid with axes $1 + h$ and $1 - h$ perpendicular to the propagation direction. The axis parallel to the propagation direction is unchanged. In addition, the quantity h changes periodically with time. The volume of the sphere does not change, so the density remains constant, i.e., unperturbed. This corresponds to the trace of the perturbation-induced, deformation tensor $\mathfrak{D}_{\alpha\beta}^{(0)}$ equaling zero in the case of a gravitational wave.

An ideal fluid, inviscid and without elasticity, is characterized by the reduction of its stress tensor $T_{\alpha\beta}$ to a single scalar—the pressure (Pascal's law). Under the action of an anisotropic deformation, as is caused by a passing gravitational wave, such a fluid responds obediently—without resistance—and takes a new shape. But this means that in the fluid, energy is not dissipated, the entropy does not grow, and new gravitational waves are not created. Now, the creation of a new wave and its interference with a passing ("old") wave are just the mechanisms for the absorption of a wave and for changing its phase. That is, on account of the radiation of the matter through which such waves pass, the index of refraction of the medium becomes different from unity. Hence, an ideal fluid does not change the propagation law for gravitational waves.

Just such a result was obtained by Lifshitz when he solved the problem of the propagation of gravitational waves in an expanding universe filled with an ideal fluid. If the fluid is viscous, however, then it heats up when deformed. The energy is evidently derived from the energy of the gravitational wave. From this Hawking (1966*a*) found an expression for the damping coefficient of a gravitational wave in a viscous medium. With $\xi = 16\pi G\eta^*/c^2$,

$$\frac{dE}{dt} = -\xi E , \quad \frac{dE}{dx} = -\frac{\xi}{c} E , \qquad (12.5.4)$$

where η^* is the viscosity. Now rewrite the viscosity in the form $\eta^* = \rho\nu$, where the kinematic viscosity $\nu \sim cl/3$ is written for relativistic particles with mean free path l. Notice that gravitational waves do not cause a change in volume or a change in temperature (in the linear approximation). Therefore, the damping of both of these types of perturbation is independent of the so-called second viscosity, i.e., independent of the processes of relaxation which tend to establish equilibrium and independent of the thermal conductivity. We then obtain $\xi = 16\pi G\rho l/3c$. It is evident, however, that such an expression is valid only until the time when one can use the notion of viscosity, i.e., as long as the mean free path l is less than the wavelength λ (or $\lambda/2\pi$). For $l \gg \lambda$, it is necessary to examine the opposite case of almost free particles interacting with a gravitational wave. We shall turn to this below. Now, without demonstrating it, we shall accept that the damping is a maximum for $l = \lambda/2 = c\tau/2$; where τ is the oscillation period of a wave. For the cosmological problem, we express the density in terms of the time, so that $\rho = 3(32\pi G t^2)^{-1}$ for $P = \epsilon/3$. We then obtain a compact expression for the upper limit on ξ:

$$\xi < \xi_{\max} = \frac{16\pi G\rho c\lambda}{6c^2} = \frac{\tau}{4t^2}. \tag{12.5.5}$$

In the course of time, the period of the wave τ grows in proportion to the wavelength according to

$$\tau \sim \lambda \sim a(t) \sim t^{1/2}.$$

Let us introduce the notation t_0 for the time when $\tau = t$, so that $\tau = (tt_0)^{1/2}$. Only after this time (t_0) can one examine the damping of a gravitational wave according to the formulae presented. It turns out that the damping integral converges:

$$\int_{t_0}^{\infty} \xi_{\max} dt = \frac{1}{4} \int_{t_0}^{\infty} \frac{(tt_0)^{1/2}}{t^2}\, dt = \tfrac{1}{2}.$$

This means that the viscosity of the medium decreases the energy of gravitational waves by no more than a factor of $e^{1/2}$ in comparison with the case in which they pass through an ideal fluid. In the latter case, only the adiabatic change of the wave energy is considered.

Let us now consider the properties of the solution (12.5.2) when $t \to 0$. During the period when the wavelength is greater than ct, this solution still does not represent a wave in the usual sense. The point is that too little time passes for even a single oscillation of the wave.

It is important that one of the independent solutions (the solution $C_5 \sin[\kappa\eta]/\eta$) is finite as $\eta \to 0 = (t \to 0)$. Consequently, if $C_4 = 0$, then bounded initial perturbations of the metric are compatible with a finite energy density ϵ_0 at earlier stages of the expansion. Indeed, the ratio ϵ_0/ϵ_r equals (in order of magnitude) the square of the dimensionless metric per-

turbation. Here ϵ_r is the energy density of electromagnetic radiation. This relation is valid for gravitational waves whose wavelength is small in comparison with the horizon size. According to this condition, however, the ratio does not depend on wavelength. It is maintained both before and after recombination, inasmuch as ϵ_r changes as a^{-4} (i.e., in the same way as ϵ_g).

12.6 Entropy Perturbations in Relativistic Theory

Entropy perturbations are of interest because they permit one to obtain the present-day structure of the Universe from a singular metric state which is accurately described (asymptotically) by a Friedmann solution and therefore exhibits no deviations from homogeneity. This circumstance was noticed by Chibisov (1972*b*).

The baryon charge, the difference $B = (n_B - n_{\bar{B}})/n_\gamma$ between the baryon and antibaryon densities, is distributed inhomogeneously in space relative to the photon density, but near the singularity the baryon charge does not influence the metric. In the limit of high temperature, moreover, the relation between the energy density and the pressure does not depend on the specific baryon charge. In particular, given a limited mass spectrum, $P = \epsilon/3$, independent of whether the density is charge-symmetric, $B = 0$, or "charged," $B \neq 0$.

A further result of Chibisov's observation is the possibility of constructing an unperturbed metric and a solution with $P = P(t)$ and $\epsilon = \epsilon(t)$ independent of the coordinates. Only during the further course of expansion, when $kT < M_n c^2$, do differences in the equation of state arise in different locations. Such differences are approximately described by the formula

$$P = \frac{\epsilon}{3}\left[1 - B(x)\epsilon^{-1/4}\right].$$

Differences in the metric and in the motion, due to differences in the equation of state, also arise. Initial entropy perturbations give rise to adiabatic density perturbations and, in particular, to growing-mode adiabatic perturbations, if the wavelength is sufficiently large. Note though that this fact is not a special consequence of the relativistic theory. It is sufficient that the perturbations be studied in an evolving Universe in which the equation of state changes in the course of time (in distinction to the Jeans formulation of the problem). Entropy perturbations corresponding to a mass between 1 $M_\odot$ and 10^4 $M_\odot$ only give rise to decaying, radiation-dominated plasma oscillations before recombination. But for masses $M > 10^5$ $M_\odot$, the entropy irregularities are preserved and are manifested after recombination. Interest in entropy perturbations has increased due to the possibility of the formation and evaporation of primordial black holes (PBH) (see chap. 24). Such a process would lead to entropy produc-

tion; perhaps it could be responsible for all the entropy! See, for example, Chapline (1976), Carr (1975, 1976), Zel'dovich and Starobinsky (1976).

In principle, one can also consider a nonuniform lepton charge distribution, i.e., the excess of neutrinos over antineutrinos. In this case, too, an unperturbed metric at the singularity can be specified. As a consequence of the large neutrino mean free path, the nonuniformity is equalized early and the amplitude of the growing-mode density perturbation turns out relatively small.

An initial perturbation in the form of a chaotic magnetic field possesses a similarity to an entropy perturbation: a magnetic field is frozen into the plasma just as is the entropy. The magnetic field, however, perturbs the metric. The magnetic-inhomogeneity hypothesis commonly proposed is rather explicit, but reasons have not yet been found for preferring it over other exotic hypotheses (see § 20.3).

12.7 The Quasi-isotropic Solution and the Hypothesis of a Uniform Distribution of Perturbations

An attractive picture of the Universe is one in which different parts have somewhat different curvature at some instant of time and, moreover, expand as different Friedmann models (but with no large local departures from these ideal models). The mathematical expression of such a picture is the quasi-isotropic solution (Lifshitz and Khalatnikov 1963*a*, *b*), which is described by the metric

$$ds^2 = c^2dt^2 - [ta_{\mu\nu}(x^\alpha) + t^2b_{\mu\nu}(x^\alpha) + \ldots]dx^\mu dx^\nu \qquad (12.7.1)$$

for the equation of state $P = \epsilon/3$.

It can be shown that this solution contains six arbitrary functions of the spatial coordinates, $a_{\mu\nu}$. Then, substituting the metric (12.7.1) into the GTR equations, one can find the $b_{\mu\nu}$ and the succeeding terms in the expansion, corresponding to the given $a_{\mu\nu}$. The specification of the $a_{\mu\nu}$ thus gives a complete set of initial conditions for a given problem. Near the singularity the main term is $ta_{\mu\nu}$. Neglecting all of the other terms, we see that each element of space expands isotropically with the same Hubble parameter in all directions. The spatial curvature, depending on $\partial^2 a_{\mu\nu}/\partial x^\tau \partial x^\lambda$, is arbitrary, but small in comparison with the $R_{0\alpha 0\beta}$ and $R_{0\alpha\beta\gamma}$ components of the spacetime curvature. In the curvature tensor, therefore, with whose help one can calculate the velocity and the distribution of the energy density of the matter, the main terms are time derivatives of $ta_{\mu\nu}$. Accounting for them gives the classical result $\rho = 3(32\pi Gt^2)^{-1}$ (for $P = \epsilon/3$). To this order, the density does not depend on the coordinates; and the matter velocity, relative to the coordinate system employed, equals zero.

An example of a quasi-isotropic model is part of a closed model "sewn"

smoothly into a flat model. We know that the asymptotic density in both parts is the same and corresponds to that given above.

Formula (12.7.1) shows that one can perform an analogous operation more generally. Namely, one can join parts of a curved space with non-isotropic curvature. It then turns out that the density is defined as

$$\rho = \frac{3}{32\pi G t^2} + \frac{f(x^\alpha)}{t} . \tag{12.7.2}$$

The function f depends on $a_{\mu\nu}$ or, more precisely, on its spatial derivatives. Evidently, then,

$$\delta = \frac{\rho - \langle\rho\rangle}{\langle\rho\rangle} = \frac{32\pi G t}{3} f(x^\alpha) \sim t . \tag{12.7.3}$$

The part of the deformation tensor with zero trace corresponds to gravitational waves of bounded amplitude. Finally, Lifshitz and Khalatnikov have shown that vortex motion is absent in this solution.

Consequently, the general quasi-isotropic solution illustrates the result, obtained by studying small perturbations, that the Friedmann behavior near the singularity ($g_{\mu\nu} \sim t$ and $\rho \sim t^{-2}$) is compatible with density perturbations and gravitational waves, but not with vortex perturbations. Furthermore, when discussing adiabatic perturbations and gravitational waves, we noticed that a finite metric perturbation must exist near the singularity if finite density perturbations and finite, nonvanishing gravitational waves are to exist today. Thus, if one asks, what cosmological solution is not in conflict with the present state of the Universe and deviates least from the strictly homogeneous solution, the answer is the *quasi-isotropic solution* (or, of course, entropy perturbations).

One can also express a hypothesis that the evolution of such a solution for reasonably selected random functions $a_{\mu\nu}(x^\alpha)$ leads to two types of equally large (in order of magnitude) metric changes. One type consists of adiabatic perturbations and the other consists of gravitational waves; initial vortex perturbations are not found in this picture. The phrase "reasonably selected" plays the role of a fig leaf concealing a theory that has not yet been worked out. Such a hypothesis, entailing equally large adiabatic perturbations and gravitational waves, was advanced by the authors (Zel'dovich and Novikov 1969) and was called the hypothesis of the "uniform distribution" of perturbations of various types.

By using this uniform distribution hypothesis, for the present epoch for gravitational waves whose wavelength is less than the present horizon size but greater than the typical distance between galaxies, the gravitational-wave energy in this region of the spectrum exceeds, by several orders of magnitude, the equilibrium energy of gravitational radiation in the same region of the spectrum if the equilibrium temperature is $T \approx 1$ K. (See chap. 17 for more detail concerning gravitational waves.) The most

important point, however, is that the equal size of the scalar, vector, and tensor metric perturbations at $t \sim t_{Pl}$ signifies that at subsequent times, when $t \gg t_{Pl}$, the vector perturbations are negligibly small in comparison with the scalar and tensor perturbations. But this means that there is no place for primordial vortex motion in the hypothesis of uniformly distributed metric perturbations. For such a possibility see Chibisov (1975).

Earlier in the exposition of the theory of small perturbations it was explained that to obtain finite density perturbations today, it is necessary to specify finite metric perturbations at the singularity. But in this situation the applicability of small-perturbation theory turns out to be uncertain. Do terms of the order of h^2 or h^3 (where h is the metric perturbation), when multiplied by functions of time, diverge as $t \to 0$? The quasi-isotropic solution completely resolves all such doubts inasmuch as in it deviations from the Friedmann solution are not presumed to be small. Moreover, those properties of the solution that we must guess at in the theory of small perturbations, are obtained in an exact, analytic form in the quasi-isotropic solution.

12.8 Long-Wavelength Perturbations and Their Representation in Terms of Spherical Waves

A study of perturbations whose wavelength is comparable with the radius of the Universe a requires mathematical methods that go beyond the simple plane-wave theory. We emphasize that a is the radius that characterizes the curvature of three-dimensional space, i.e., the quantity defined for a closed or for a hyperbolic model and not the distance to the horizon. To exclude the flat model, we assume $\Omega_0 \neq 1$. The ratio of the wavelength of the perturbation to the radius of the model is evidently constant. It does not change during the expansion.

For a closed model the solution is given in the classic work by Lifshitz (1946). The characteristic functions into which one should decompose perturbations have a discrete spectrum and are constructed by analogy with the method used for constructing ordinary spherical harmonics. We shall henceforth confine ourselves to the case of a scalar function, which describes, as we know, adiabatic density perturbations.

The function $\psi = \psi(r, \theta, \phi)$, the function of the spatial, comoving coordinates that describes the scalar perturbations, satisfies the equation

$$\nabla^2\psi = \frac{1}{\sin^2 r}\frac{\partial}{\partial r}\left(\sin^2 r \frac{\partial\psi}{\partial r}\right) + \frac{1}{\sin^2 r} \times \left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}\right] = -n(n+2)\psi\,. \tag{12.8.1}$$

The boundary condition on ψ is that it must be bounded at $r = 0$ and at $r = \pi$; furthermore, $n = 0, 1, 2, \ldots$.

The solution of equation (12.8.1) is characterized by the three indices n, l, m, and can be written in the form of a product of three functions $R_{nl}(r)P_l^m(\theta)\Phi_m(\phi)$. The dependence on ϕ is given by sin $m\phi$ or cos $m\phi$, so that m is the number of nodal surfaces (places where $\psi = 0$), $\phi = \text{const}$. The dependence on θ is given by the Legendre polynomials P_l^m (cos θ) with $l - m$ the number of nodes. Finally, the dependence on r is given by the functions $R_{nl}(r)$, where $n > l$. Near the origin, this function is proportional to r^l; in particular, only the function R_{n0} differs from zero at $r = 0$. The number of radial nodes equals $n - l$.

The case of the open (hyperbolic) model is in many respects analogous. The equation for the function ψ takes the form

$$\nabla^2\psi = \frac{1}{\sinh^2 r}\frac{\partial}{\partial r}\left(\sinh^2 r\,\frac{\partial\psi}{\partial r}\right) + \frac{1}{\sinh^2 r}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}\right] = (k^2 + 1)\psi\,. \tag{12.8.2}$$

The variables are separable so that the dependence on θ and ϕ is again given by the functions P_l^m (cos θ) sin $m\phi$ or P_l^m (cos θ) cos $m\phi$. After substituting these, the equation

$$\frac{1}{\sinh^2 r}\frac{d}{dr}\left[\sinh^2 r\,\frac{dR}{dr}\right] - \left[1 + k^2 + \frac{l(l+1)}{\sinh^2 r}\right]R = 0 \tag{12.8.3}$$

is obtained for the function R_{kl} (r). One property of these radial functions is that

$$R_{kl}(r) \sim r^l \quad \text{for} \quad kr \ll 1\,. \tag{12.8.4}$$

Furthermore, the asymptotic behavior differs depending upon whether $k > 1$ or $k < 1$. In the first case, there is flat, intermediate asymptotic behavior, $R \sim r^{-1}\sin kr$, in the region with $kr > l$ and $r < 1$. In the region $r > 1$, $kr > l$, the asymptotic behavior, $R \sim \sin kr/\sinh r$, is connected in an essential way with the hyperbolic metric of the model. That is, it is related to the facts that the surface area of a sphere in such a model grows as $4\pi \sinh^2 r$, and that

$$4\pi \sinh^2 r \gg 4\pi r^2\,, \tag{12.8.5}$$

for large radii. Inasmuch as r is bounded only from below, $r \geq 0$, any real k is possible for any integer l, so the spectrum is continuous. This is certainly natural for an infinite volume. The functions $\psi \sim \sin kr/\sinh r$ with k real vary fast enough on scales greater than a. For functions which vary slowly on scales greater than a, it is necessary that k be complex: $k = \gamma + i\delta$, with $|\delta| < 1$ (Lukash 1975). The quantity δ is constant, and γ varies. The physical meaning of δ is the following. For $r \to \infty$, the asymptotic behavior is $R_{kl}(r) \sim \sin(\gamma r + \phi_l)\exp[-(1 - |\delta|)r]$. Thus,

$|\delta|$ determines the correlation scale, r_k, of the perturbation; for $r_k = \infty$, $|\delta| = 1$. We now return to the $\delta = 0$ case.

It is easy to express the concepts of statistical homogeneity and isotropy of flat space using a system of plane waves.

Plane waves with a given k^2 but with different directions are equally likely in a statistical sense. It then follows that spherical waves normalized to the same asymptotic behavior $2^{1/2} \sin(kr + \phi_l)/(r\pi^{1/2})$ for $r \to \infty$ are also equally likely for the same k, independent of l. Hence, density perturbations can be written as

$$\begin{aligned}\frac{\delta\rho}{\langle\rho\rangle} = &\int dk C_{k00} R_{k0}(r) P_0^0(\theta) \\ &+ \int dk C_{k10} R_{k1}(r) P_1^0(\theta) + \int dk C_{k11,c} R_{k1}(r) P_1^1(\theta) \cos\phi \\ &+ \int dk C_{k11,s} R_{k1}(r) P_1^1(\theta) \sin\phi \\ &+ \int dk C_{k22,s} R_{k2}(r) P_2^2(\theta) \sin 2\phi + \ldots .\end{aligned} \tag{12.8.6}$$

The functions C_{klm} which depend on the scalar variable k, the modulus of the wave vector, are random functions. The second and third indices characterize the total angular momentum and its projection.

Homogeneity and isotropy for the Universe, including statistical homogeneity and isotropy for the perturbations, means that the $C_k \ldots$ with diverse indices l, m, and s or c are the same; that they have the same mean square values and are independent, i.e., uncorrelated with one another.

A density perturbation at the point of observation (at the origin of coordinates, $r = 0$) depends only on the first integral in formula (12.8.6), i.e., on C_{k00}. This is because only $R_{k0}(r)$ is nonvanishing at $r = 0$. Thus, $\delta\rho(0) = \langle\rho\rangle \int dk C_{k00}\, k(\pi\sqrt{2})^{-1}$.

Motion of the matter located at the origin of coordinates relative to the RR is written in terms of the velocity components along the three coordinate axes. Such motion depends on the following three integrals only:

$$\begin{aligned} u_x(0) &= H \int dk C_{k11,c} (\pi\sqrt{6})^{-1} ; \\ u_y(0) &= H \int dk C_{k11,s} (\pi\sqrt{6})^{-1} ; \\ u_z(0) &= H \int dk C_{k10} (\pi\sqrt{6})^{-1} . \end{aligned} \tag{12.8.7}$$

These three components are given for the flat model with $\Omega_0 = 1$ and $H = 2/(3t)$.

The statistical mean values of these quantities are given by the expressions

$$\langle(\delta\rho/\rho)^2\rangle = \int \langle |C_k|^2 \rangle \frac{k^2}{2\pi^2} dk ; \tag{12.8.8}$$

$$\langle u^2 \rangle = H^2 \int \langle |C_k|^2 \rangle \frac{dk}{2\pi^2} . \qquad (12.8.9)$$

Here $\langle |C_k|^2 \rangle$ is the mean square value of the random function taken over all of the different l and m.

One cannot take the position of an observer as random, for the formation of the Galaxy, of the Sun, and of civilization is most probable where $\delta\rho > 0$. But from the decomposition we see that for the linear case that the average velocity is independent of the local density.

The question of the periodic distribution of any particular kind of celestial object is most conveniently considered with the help of spherical functions. G. Burbidge has carried out persistent searches for such regularities. Chapter 4, on the cosmological constant, describes the birth and death of the hypothesis that quasars are concentrated at the redshift $z = 1.95$. Burbidge and O'Dell (1972) have established with the help of Fourier analysis that there are periods $\Delta z = 0.031$ and $\Delta z = 0.062$ in the distribution of quasars with respect to the redshift. We shall not discuss the statistical significance of these conclusions here, however. We shall, instead, turn to the following questions: What kind of periodicity is compatible with the assumption of a cosmological origin for the redshift of quasars? Assume that $\Omega_0 = 1$ so that the Universe is flat. In this case the redshift is related to the distance, in terms of the comoving coordinate r, by the formula

$$r = \int \frac{dt}{b(t)} = \frac{1}{b_0 H_0} \int_0^z \frac{dz}{(1+z)^{3/2}} = \frac{2}{b_0 H_0} [1 - (1+z)^{-1/2}] . \qquad (12.8.10)$$

The calculated number of quasars, over the whole sky (integrated with respect to θ and ϕ), depends only on the zero-order ($l = 0$) harmonic. The most extreme assumption, that corresponding to periodicity, is that $C_k = A\delta(k - k_0)$. This leads to the distribution

$$n = \langle n \rangle \left(1 + \alpha \frac{\sin \psi}{\psi}\right), \quad \psi = \beta[1 - (1+z)^{-1/2}] . \qquad (12.8.11)$$

The maximum value of α, for which n is everywhere positive, is $\alpha \approx 4.5$. Requiring further that the period equal $\Delta z \approx 0.062$ for small z, we then obtain $\beta = 200$. Given this, then for $z \approx 2$, the period Δz is about 0.3 and the amplitude $|(n - \langle n \rangle)/\langle n \rangle|_{\max}$ is only 0.05. Such behavior differs substantially from the simple periodic dependence suggested by Burbidge and O'Dell. The expansion of the perturbations in terms of spherical harmonics thus suggests another method for interpreting observations. It is also not difficult to derive similar formulae applicable to the hyperbolic model.

In the expression for the velocity of an observer relative to the RR, a simplification was made. It was assumed that the perturbations represent

growing modes of matter motion on a background of radiation at rest. Such an approach is entirely warranted in the case of distances that are small compared with the horizon. Long waves, however, draw the radiation into motion along with the matter. At the same time as u decreases in the limit $\lambda \to \infty$, the velocity of the matter relative to the radiation tends to zero. The matter velocity is observable as a dipole component in the temperature deviations $\Delta T/T$. A detailed calculation of this effect, especially for the open model ($\Omega_0 < 1$), requires the use of the spherical harmonic method.

One indeed ought to expect that the spherical-wave method will find wide application in the theory of perturbations of the homogeneous, isotropic Universe in the very near future.

13 STATISTICAL THEORY

13.1 Randomness and Fourier Analysis

In the preceding sections we examined, in great detail, the behavior of a discrete perturbation which depends periodically on the spatial coordinates (a "plane wave"). This type of perturbation was considered because of the mathematical simplification that results. The problem reduces to the study of the time dependence of the amplitude and the phase. The spatial distribution of a plane wave perturbation evolves such that the wavelength grows in proportion to the scale factor while the amplitude depends on time in a manner studied above.

Now, as long as the perturbations are small, any perturbation (arbitrarily distributed in space) can be represented as a linear combination of plane waves each of which acts independently. With this in mind, we shall, in the subsequent development, study the laws governing perturbations of arbitrary form. For perturbations with a small (but not too small) amplitude, one can push the theory one or two steps further by examining as corrections to the linear theory the mutual interaction of waves. In several cases, one can successfully solve the complete nonlinear problem either exactly or approximately.

A characteristic peculiarity of the cosmological problem is the chaotic character of the perturbations. We shall begin our analysis with a study of this point.

Even a fleeting glance at the distribution of galaxies shows that their location and form is, to a large degree, random. Clearly, then, statistical methods are necessary for a description of the structure of the Universe on the scales of galaxies and clusters of galaxies. But statistical methods are also (or must also be) an important part of the evolutionary theory, particularly of the theory describing the evolution of perturbations. One would like, therefore, by taking random initial perturbations, to obtain a statistical description of the observable Universe by applying to the perturbations the fundamental laws (of gravitation, of interactions with radiation, and so forth).

The basic propositions of the statistical description, of the statistical laws, and of the theory of probability are no doubt insufficiently familiar to a majority of readers and therefore require some explanation. In the

simplest form, for example, in connection with the tossing of coins or dice, randomness is defined as the opposite of any functional relationship. The result of each successive event must not depend on past events (if the game is honest!). Now the first difficulty with our problem is that we are not working with discrete experiments and quantities (separate tosses of the coin with the answers "yes" or "no"), but with functions of continuous variables. An example is the density, which is a function of the spatial coordinates and of the time. With the help of an artificial division of space into separate boxes, however, one can introduce the discrete set $\{\langle \rho_i \rangle\}$, where $\langle \rho_i \rangle$ is the average density in the box with the index i. Assume, then, that the $\langle \rho_i \rangle$ and $\langle \rho_k \rangle$ and so forth are independent at some initial time. Very quickly this independence is destroyed by the physical interaction between the matter in neighboring boxes. These considerations show that the box approximation is not suitable. Even the form of the description itself must be adapted to the character of the processes occurring in the systems studied below.

When a homogeneous distribution of matter is a reasonable first approximation (i.e., the perturbations are small), plane waves are naturally distinguished by means of simple properties. In the linear approximation, plane waves change independently of one another, i.e., without interacting. Under this circumstance, it is advisable to make use of a Fourier decomposition of the random functions.

Great successes have been achieved by the use of random variables and their Fourier expansions in connection with the theory of turbulence, plasma theory, and in radio engineering. The best work on this subject, as applied to hydrodynamics, is the remarkable book by Monin and Yaglom (1965, 1967).

The Fourier expansion of a function given in a volume V can be written in the form

$$\delta(x, y, z) \equiv \frac{\rho(x, y, z) - \langle \rho \rangle}{\langle \rho \rangle} = V^{-1/2} \sum_k A_k e^{i\mathbf{k}\cdot\mathbf{x}}, \quad (13.1.1a)$$

where

$$A_k = V^{-1/2} \int_V \left[\frac{\rho(x, y, z) - \langle \rho \rangle}{\langle \rho \rangle}\right] e^{-i\mathbf{k}\cdot\mathbf{x}} d^3x. \quad (13.1.1b)$$

From (13.1.1b) it is clear that A_k has the property: $A_k = A^*_{-k}$. Thus, although the right-hand side of (13.1.1a) is written in complex form, it is real. When we speak below of the independence of A_k with respect to different $\mathbf{k}$, keep in mind the rigid connection between the pairs A_k and $A_{k'}$ for $\mathbf{k}' = -\mathbf{k}$. Notice that the expansion is carried out with respect to a system of orthonormal functions:

$$\delta = \sum_k A_k \psi_k(x), \quad \psi_k(x) = V^{-1/2} e^{i\mathbf{k}\cdot\mathbf{x}}.$$

The orthonormality conditions are

$$\int_V \psi_k \psi_{k'} d^3x = \delta_{k,k'} ; \quad \delta_{k,k'} = \begin{cases} 0 , & k \neq k' \\ 1 , & k = k' . \end{cases}$$

The $\{k\}$ are a discrete set of three-dimensional vectors. From the normalization condition it follows that the functions ψ_k have the dimensions $V^{-1/2} \leftrightarrows$ $\mathrm{cm}^{-3/2}$. Since $\delta = (\rho - \langle\rho\rangle)/\langle\rho\rangle$ is dimensionless, the coefficients A_k evidently have the dimensions $\mathrm{cm}^{3/2}$. We hope that there will be no confusion between the symbols for density perturbations and those for the Dirac or Kronecker delta functions.

It would be convenient for us to consider coefficients without dimensions; even knowing the numerical value of the A_k, one could not say if the perturbations are large or small. Moreover, we would also like to work with a quantity that takes account of the statistical character of the problem. In the subsequent development, therefore, we shall introduce a dimensionless quantity proportional to the square of the amplitude, $|A_k|^2$, and to the cube of the modulus of the wave vector, averaged to remove the statistical variations:

$$\Delta_k = \frac{k^3}{2\pi^2 n} \sum |A_k|^2 .$$

Here n ($\gg 1$) is the number of terms in the summation. The quantity Δ is dimensionless because, while the dimensions of $|A_k|^2$ are cm^3, those of k are cm^{-1}, so those of k^3 are cm^{-3}. The factor $2\pi^2$ is introduced for convenience. If, for all wave vectors, the dimensionless quantity Δ_k is small, then the density perturbation ($\delta = \delta\rho/\rho$) is small. The proof of this will be given below.

The wave vector k can assume one of the discrete values permitted for the given, bounded volume V. For example, if the volume is cubic, then, with $V = L^3$,

$$k_x = 2\pi n_x/L , \quad k_y = 2\pi n_y/L , \quad k_z = 2\pi n_z/L , \qquad (13.1.2)$$

where n_x, n_y, and n_z are whole numbers. For definiteness, we exclude the constant term ($n_x = n_y = n_z = 0$) from the summation. Here this is justified because, by definition, $\langle\delta\rangle_V = 0$, if by $\langle\rho\rangle$ we mean the average over the given volume V.

A reasonable definition of a random function is a function whose Fourier coefficients are random. The idea is that this randomness is not resolved by the physical interaction for at least as long as the perturbation is small. Consequently, the assumption of the randomness of the Fourier components A_k is better than an assumption of the randomness of values $\delta(x)$ in various boxes.

The randomness hypothesis is connected with the idea that we can choose many different volumes in the Universe. Each of them is called a realization and has a definite density $\rho = \langle\rho\rangle [1 + \delta (x, y, z)]$ and a single set of A_k.

Examining these volumes jointly, we can ask: How often is a given value A_k encountered? In answering this question, we can introduce the probability $P(A_k)$ or, for many waves jointly, the probability $P(A_{k_1}, A_{k_2})$, for the appearance of given values of the Fourier coefficients. In the linear theory, it is convenient to use the complex notation $A_k = B_k + iC_k$, where the B_k and C_k are real. Then it is simpler to discuss the probability $P(B'_k, C'_k, B''_k, C''_k, \ldots)$. If one takes N volumes, then the number of volumes dN in which B'_k lies between $B_k^{(0)}$ and $B_k^{(0)} + dB'_k$ is given by

$$dN = NP(B_k^{(0)})dB'_k . \tag{13.1.3}$$

Concerning the random distribution of the quantity of interest to us, it is natural to assume that $P(B'_k, C'_k; B''_k, C''_k)$ is equal to the product of terms of the type

$$\frac{1}{2\pi\beta'_k\gamma'_k} \exp\left(-\frac{B'^2_k}{2\beta'^2_k}\right) \exp\left(-\frac{C'^2_k}{2\gamma'^2_k}\right)$$

with specified coefficients β'_k and γ'_k. Presenting P in the form of a product of factors depending only on the $B'_k, C'_k, \ldots$ (a factorization) signifies that the $B'_k, C'_k, \ldots$ are independent. The selection of a Gaussian form for the factors signifies that we have selected a normal distribution law. Though this is not the most general distribution, it is as a rule realized in practice.

In the theory of probability it is proved that if a quantity, A_k, is the sum of a large number of independent quantities, a_{kn}, then for any likely distribution law for each constituent a_{kn}, a normal law is obtained for the sum

$$A_k = \sum_n a_{kn}$$

(in the limit of large n). At an early stage near the singularity, the integral which determines the Fourier coefficients reduces to a sum over causally disconnected regions (if there is no period "before the singularity"). Hence, the assumption of a normal distribution for the A_k is natural.

Another possible form for the distribution is $A_k = a_k \exp(i\phi_k)$. This is described by specifying $P(a'_k, \phi'_k; a''_k, \phi''_k; \ldots)$. Instead of taking the real part of a complex expression, one can then make use of the condition $A^*_{-k} = A_k$, which also guarantees reality. The normal distribution law in this case signifies that P depends only on the a_k, and not on the ϕ_k for which all values are in the range $0 \leq \phi_k \leq 2\pi$, so that $P \sim (1/\alpha_k) \exp(-a_k^2/2\alpha_k^2)$.

The parameters β_k, γ_k, and α_k entering into the definition of P determine the average squares of the B_k, C_k, and $a_k = A_k$. That is, it is also possible to formulate our assumptions in terms of the β_k, γ_k, and α_k. For example, the isotropy and homogeneity of the model are expressed in this language by saying that the β_k, γ_k, and α_k depend only on the modulus of $\boldsymbol{k}$, not on its direction, and, in addition, that $\beta_k = \gamma_k$ and $\alpha_k = \beta_k\sqrt{2}$.

13.2 The Correlation Function and the Sizes of Self-gravitating Objects

Though a Fourier expansion is very convenient for dynamical calculations, the function $\beta(k, t)$ is not the kind of answer that warms the hearts of astronomers. An astronomer has the right to ask what the number and the forms are of clusters and of other objects. Here though, the theorist must offer partial apologies: an exact theory of cluster formation requires complicated nonlinear calculations which are very difficult to perform.

If, however, the astronomer does not lose interest in theory at this point, and he asks about density inhomogeneities not exceeding the average density by more than 10–20%, the linear theory applies. But even for these (significantly less interesting) inhomogeneities the astronomer wants to know their form and their amplitude—not the rather abstract function $\beta(k)$.

In this way the problem arises of making the passage back from the Fourier language to one describing the spatial distribution of discrete objects. For a single realization, the answer is quickly given (see eqs. [13.1.1]). But in a statistical approach to the question, a particular function $\rho(x, y, z)$ has no meaning. We can only speak about average quantities—ones averaged over a large number of realizations. Evidently $\langle \delta(x, y, z) \rangle = 0$, i.e., the average value of $\delta(x, y, z)$ vanishes at each point in space. There is meaning, however, in talking about the average of the square of the density perturbation, $\langle \delta^2(x, y, z) \rangle$, which is not zero. Making use of the properties of the Fourier series, it is easy to show that

$$\langle \delta^2(x, y, z) \rangle = \frac{1}{V} \sum_k \langle A_k^2 \rangle . \tag{13.2.1}$$

In performing the sum we must recall equations (13.1.2) which determine the set of values of $\{\boldsymbol{k}\}$. In $\boldsymbol{k}$-space the points with integer values of n_x, n_y, n_z are located in a cubic lattice; in this space their density is $\sigma = (L/2\pi)^3 = V/8\pi^3$. The number of independent waves in the interval δk around k is equal to the product of the density and the volume of the corresponding spherical shell in k-space:

$$dN = \sigma \times 4\pi k^2 \delta k = \frac{L^3 k^2 \delta k}{2\pi^2} = \frac{V}{2\pi^2} k^2 dk . \tag{13.2.2}$$

Therefore,

$$\langle \delta^2(x, y, z) \rangle = \frac{1}{2\pi^2} \int \langle A_k^2 \rangle k^2 dk = \sum \Delta_k \frac{\delta k}{k} . \tag{13.2.3}$$

Since the A_k and Δ_k remain finite as $V \to \infty$, the sum may be replaced by an integral; thus

$$\Delta \equiv \langle \delta^2(x, y, z) \rangle = \int_0^\infty \Delta_k \frac{dk}{k} ; \qquad \Delta_k \equiv \frac{\langle A_k^2 \rangle k^3}{2\pi^2} . \tag{13.2.4}$$

Given now the normal (Gaussian) distribution law for B_k and C_k, the distribution law for δ is also the normal law:

$$P(\delta) \quad \frac{1}{(2\pi\Delta)^{1/2}} \exp\left(-\delta^2/2\Delta\right) . \tag{13.2.5}$$

The left hand side is the probability that a given value of δ is obtained at any given point. The quantity Δ is evidently the same for all points in space and does not depend on time.

These formulae answer the question concerning the amplitude of the inhomogeneities, but say nothing about their spatial structure. It is rather the correlation function $f(r)$ that characterizes the spatial structure of the inhomogeneities. This function is defined by

$$f(r) \equiv \frac{\langle \delta(x)\delta(y)\rangle}{\langle \delta^2\rangle}, \quad r = x - y . \tag{13.2.6}$$

What is the meaning of the correlation function? Imagine a space marked off into regions of average size r_0 with $\delta > 0$ and similar regions with $\delta < 0$. Select a point x inside of a region in which $\delta > 0$, so that $\delta(x) > 0$. Upon moving away from this point by a distance less than r_0, we are very likely to remain in a region with $\delta > 0$. Therefore, $\langle \delta(y)\rangle > 0$ if $|x - y| < r_0$ and $\delta(x) > 0$.

With a probability greater than one-half, the product $\delta(x)\delta(y)$ is positive if $|x - y| < r_0$. The same assertion also holds for the average. By virtue of the quadratic nature of $f(r)$, then, the regions where δ is negative give a contribution of the same sign—positive—to $f(r)$. Thus $f(r)$ is positive for $r < r_0$.

We define δ in such a way that $\langle \delta\rangle = \int \delta d^3x = 0$. From this it follows that the average value of the correlation function over volume is also zero. To see this, write $r = y - x$, or $y = x + r$, so that

$$\int f(r)d^3r = \frac{1}{\langle \delta^2\rangle} \int \left\{\delta(x)\left[\int \delta(x + r)d^3r\right]\right\} d^3x .$$

The second integral (the one within the square brackets) vanishes because it is taken over the whole volume, and in this case the choice of the origin plays no role. Hence,

$$\int f(r)d^3r \equiv 0 .$$

In a Universe that is homogeneous and isotropic on the average, the correlation function depends only on the absolute magnitude of the distance, $r = |r|$, and not on the direction of the vector r. We can therefore rewrite the above equation as (omitting a factor of 4π)

$$\int_0^\infty f(r)r^2dr = 0 .$$

Thus, $f(r)$ must necessarily be a function whose sign varies. By definition, $f(0) = 1$ (because of the choice of denominator in eqs. [13.2.6]). Further, $f(r) > 0$ for $r < r_0$. On the other hand, $f(r)$ must be negative somewhere.

The natural conclusion is that the first zero of the function f, i.e., the smallest value of r, r_1, for which $f(r) = 0$, gives the average size of regions with the same sign of δ; i.e., $r_1 \approx r_0$.

We shall not hide from the reader that the choice of definition of the correlation function is dictated by the requirements that it be both easy to compute and easy to interpret physically. With this in mind, we shall give an expression for $f(r)$ in terms of the spectral function $\beta(k)$ characterizing the amplitude of waves of various length. This expression is

$$f(r) = \frac{\int_0^\infty \beta^2(k) \frac{\sin kr}{kr} k^2 dk}{\int_0^\infty \beta^2(k) k^2 dk} = \frac{\int_0^\infty \Delta_k \frac{\sin kr}{kr} \frac{dk}{k}}{\int_0^\infty \Delta_k dk/k} . \quad (13.2.7)$$

Knowing $f(r)$, one can then find r_0. If, for example, $\beta^2(k)$ is concentrated in a narrow interval around k_0, then, as is apparent from equation (13.2.7) $r_0 \approx \pi/k_0$; r_0 is then equal to half a wavelength.

One can now express the assumption relating to the time when a significant fraction of the mass passes into gravitationally bound objects and relating to the masses of such objects. We suppose that most $\beta(k)$ (those with $k < k_J$; k_J is the Jeans wavenumber) grow with time as a consequence of gravitational instability. The quantity Δ also grows. When, in the linear theory, Δ is of order unity, then one can consider, roughly, that fragmentation into discrete parts has taken place.

Let us now write the expression for δ^2 in the form

$$\Delta = \langle \delta^2 \rangle = \frac{1}{(2\pi)^3} \int_0^\infty [4\pi\beta^2(k)k^3] d \ln k = \int_0^\infty \Delta_k d \ln k . \quad (13.2.8)$$

We suppose every function $\beta(k)$—or a large part of them (those with $k < k_J$, with k_J the Jeans critical value)—grows as time passes as a consequence of gravitational instability. The quantity Δ correspondingly grows. When, in the linear theory, Δ achieves the order of unity, then one can consider roughly that to the first approximation the process of the disintegration of matter into discrete parts has taken place. The function $\Delta_k = 4\pi\beta^2 k^3/(2\pi)^3$ introduced above (with $\langle A_k^2 \rangle = \beta^2$) in this formula is dimensionless. If in this function Δ_k there is a broad maximum at some value k_m (corresponding to an interval in $\ln k$ of order unity), then, in order of magnitude,

$$\Delta = \langle \delta^2 \rangle \approx \Delta_k(k_m) . \quad (13.2.9)$$

It is even better to pass from k to the corresponding mass, $M \equiv \rho(\pi/k)^3$ and to define $\Delta(M) \equiv \frac{1}{3}\Delta_k$ so that

$$\Delta = \int_0^\infty \Delta(M) d \ln M . \quad (13.2.10)$$

The first zero, $r = r_1$, of the correlation function is roughly determined by $\sin k_m r_1 = 0$, so $r_1 = \pi/k_m$. This corresponds, as before, to half a wavelength.

We note here that Peebles and Hauser (1973) prefer the covariance function, $W(\theta)$, defined by the excess probability that an object will be found within a solid angle $d\Omega$ centered at an angular distance θ from a randomly chosen object.

$$dP = \mathfrak{N}[1 + W(\theta)]d\Omega\,, \tag{13.2.11}$$

where $\mathfrak{N}$ is the average number of objects per unit solid angle. Obviously, if the distances (and not only the angular positions) were known, the corresponding spatial covariance function, $W(r)$, could be calculated. The difference between the *correlation* function $f(r)$ and the *covariance* function $W(r)$ is the different normalization. For a continuous matter distribution, the spatial covariance function is

$$\begin{aligned} W(r) &= \langle[1 + \delta(x)][1 + \delta(x + r)]\rangle - 1 = \langle\delta(x)\delta(x + r)\rangle \\ &= \langle\delta^2\rangle f(r)\,. \end{aligned} \tag{13.2.12}$$

Thus, since $W(r = 0) = \langle\delta^2\rangle$, the covariance function gives more information than the correlation function [for which $f(r = 0) \equiv 1$]. The covariance function is also more convenient when considering discrete objects.

We conclude with a note of caution. For a broad spectrum of perturbations, the zero of the correlation function may be very large ($r_1 \to \infty$). A better condition for estimating the scale on which objects have separated out from the expanding background is $W(r) \approx 1$.

13.3 Deviations from the Average Density in a Given Volume

A natural path to a description of an inhomogeneous matter distribution consists in taking a sphere of radius r and determining the mass contained in this sphere. This mass, of course, depends on the choice of the center of the sphere. In statistical problems, however, we can find the average mass $\langle M\rangle$ and the average size of a dimensionless deviation from the average,

$$\mu \equiv \frac{\langle\Delta M\rangle}{\langle M\rangle} = \frac{(\langle M^2\rangle - \langle M\rangle^2)^{1/2}}{\langle M\rangle}\,. \tag{13.3.1}$$

This quantity, considered as a function of the radius, or of the mass itself (i.e., $\mu[\langle M\rangle]$), is apparently a good or, at least, the simplest characteristic of inhomogeneity in a matter distribution.

Imagine that matter is distributed in the form of isolated bodies each of mass M_1. Let the average density of a body be ρ_1. The number of bodies per unit volume is then $n = \langle\rho\rangle/M_1$. Thus, each body is associated with a volume $V \approx n^{-1} = M_1/\langle\rho\rangle$; at the same time, the volume of a body itself is $v = M_1/\rho_1$. We assume, for simplicity, that all of the matter is congregated in isolated bodies, so that no matter remains between the bodies. In

this case it is clear that the function $\mu(\langle M\rangle)$ behaves in the following manner:

1. For $\langle M\rangle \gg M_1$, many $(\langle M\rangle/M_1)$ bodies are found in the volume under consideration. Thus, μ is small; $\mu \lesssim 1$ for $\langle M\rangle \gtrsim M_1$ and $\mu \to 0$ for $\langle M\rangle \to \infty$.

2. If $(\langle\rho\rangle/\rho_1)M_1 < \langle M\rangle < M_1$, then sometimes there is only one body and sometimes not even one in the volume under consideration. One can show that in this region

$$\mu \approx (M_1/\langle M\rangle)^{1/2} > 1 .$$

3. Finally, for $\langle M\rangle < \langle\rho\rangle M_1/\rho_1$, i.e., for $\langle M\rangle < \langle\rho\rangle v$, the volume under consideration is less than the volume of a single body. At the same time, if there is no structure within the body (constant density ρ_1 inside the body), we find that μ does not depend on $\langle M\rangle$ and does not grow with any further decrease in $\langle M\rangle$. Under this circumstance,

$$\mu \approx (\rho_1/\langle\rho\rangle)^{1/2} > 1 .$$

The resultant curve $\mu(\langle M\rangle)$ allows one to obtain an idea of the matter distribution. For example, an inhomogeneous matter distribution inside of discrete bodies would lead to a further rise in the curve in a region of still smaller values of $\langle M\rangle$. However, this structure does not influence the part of the curve for $\langle M\rangle > \langle\rho\rangle v$. This example shows how tempting it is to make use of the function $\mu(\langle M\rangle)$ to systematize the observations and to facilitate comparison of them with theory.

If a matter distribution is specified by the coefficients of a Fourier series, then $\mu(\langle M\rangle)$ is determined by the formula

$$\mu^2(\langle M\rangle) = \frac{9}{4\pi}\int_0^\infty \beta^2(k)(kr)^{-3}J_{3/2}{}^2(kr)k^2dk , \qquad (13.3.2)$$

where $\langle M\rangle = 4\pi\langle\rho\rangle r^3/3$ and $J_{3/2}(x)$ is a Bessel function.

Proceeding further, we shall, as before, be considering the case of matter concentrated in isolated bodies of mass M_1. How does the function $\mu(\langle M\rangle)$ behave for $\langle M\rangle \gg M_1$, i.e., for objects containing many of the smaller bodies? An analysis of this case shows (see below) that a substantial improvement in the definition of $\mu(\langle M\rangle)$ is possible. A naive answer to this question is that discrete objects behave much as does a statistically independent distribution of particles. The average number of little bodies in such an object is $\langle N\rangle = \langle M\rangle/M_1$, while

$$\mu = \frac{\langle\Delta N\rangle}{\langle N\rangle} = \frac{1}{\sqrt{N}} = \left(\frac{M_1}{\langle M\rangle}\right)^{1/2} < 1 \quad \text{for} \quad \langle M\rangle > M_1 . \quad (13.3.3)$$

But this result is not always correct! In reality, there is no universal answer, since the behavior of the curve in the region where $\mu < 1$ depends on the isolation process. Indeed, a study of the processes in large volumes con-

taining many objects and leading to small values of μ can give valuable information about the Universe and its large scale structure.

The law $\Delta N = \sqrt{N}$ is only obtained under the assumption of a random not a correlated—arrangement of discrete objects in space. One can thus say that this law corresponds to the hypothesis of a God who, from outside, sows space with galaxies. It is assumed that the probability that each succeeding galaxy falls in one or another region of space is the same at all points, independent of how the preceding galaxies are distributed. But such a hypothesis is evidently unacceptable.

A natural formulation of the question consists, rather, of an evolutionary examination of the formation of the objects studied (galaxies or clusters of galaxies). It is here assumed that in the beginning, a plasma is distributed almost uniformly and that galaxy formation is a result of the gravitational-instability-caused growth of small perturbations.

At first glance such a hypothesis leads to the quite definite conclusion that ΔN must be less than $\sqrt{N}$; less than in the hypothesis of "randomness." Now, in the process of perturbation growth, the density increase at any place and the formation of gravitationally bound objects, are results of the inflow of matter from neighboring regions. It would therefore seem, following from this, that the probability for the formation of another galaxy next to the given one is less than for the formation of one farther away. Such a negative correlation must lead to a decrease in the fluctuations in large regions with a large value of $\langle N \rangle$. This simple reasoning, however, valid in the case of the formation of droplets of mist in supercooled vapor, is not valid in the case of gravitational instability. The reason is that gravity is a long-range force.

To see this in more detail, imagine that an excess density is concentrated in a small volume. The gravitational field of the excess mass is characterized by a potential decreasing with distance: $\delta\phi = -G\delta M/r$. This corresponds to an acceleration equal to $-G\delta M/r^2$ and directed toward the center of mass. The motion caused by this acceleration is characterized by a radial velocity

$$u_r = -\frac{G}{r^2} \int \delta M(t) dt \, .$$

The flux of matter transported inside a sphere of radius r is hence

$$q = 4\pi r^2 \rho |u_r| = 4\pi\rho G \int \delta M(t) dt \, .$$

From this one can construct the equation

$$\frac{d}{dt} \delta M = 4\pi G\rho \int \delta M(t) dt \, ,$$

the solution to which coincides with that of the Jeans theory (see chap. 10):

$$\delta M_i = \text{const.} \exp\left[(4\pi G\rho)^{1/2} t\right] ;$$

$$\delta M_d = \text{const.} \exp |-(4\pi G\rho)^{1/2} t| .$$

Here, however, our goal is to emphasize the $u \sim r^{-2}$ dependence, a dependence according to which the amount of matter in each intermediate layer does not change (the flux q does not depend on r: $\nabla \cdot u = 0$). One can say that the increase in mass at the center occurs because of the inflow of mass from infinity, and not from the neighboring regions! Hence, for gravitational instability, there is no anticorrelation among neighboring galaxies. Recall too that it was shown in § 10.3 that small perturbations grow in a similar way.

The final conclusion is that the fluctuation law $\Delta N = f(N)$ for the distribution of discrete objects depends on the law for the original small perturbations. In principle, a law such as $\Delta N \sim N^{-1/6}$ or $\Delta N \sim N^{2/3}$ is possible, depending on the spectrum of small perturbations. Observational studies of fluctuations can thus give information concerning the initial state.

It is instructive to analyze the situation with the help of a Fourier spectrum. A single object of volume V and linear size $l \approx V^{1/3}$ is described by the condition $A_k = \text{const.}$ for $kl < 1$; i.e., the A_k do not depend on k. Moreover, randomly distributed objects also are described by the spectrum $\beta_k = \text{const.}$ The individual A_{k_1}, A_{k_2}, etc., have their phases randomly distributed. Therefore, in calculating the overall $A_k = A_{k_1} + A_{k_2} + \ldots$ we shall obtain different values for nearby values of k. On the average, the A_k are statistically distributed: $\langle A_{k_i} A_{k_j} \rangle = 0$ and therefore $\langle |A_k|^2 \rangle = 2\beta_k^2$. A flat spectrum for each A_{k_i} leads to a statistically flat overall spectrum. If there is an anticorrelation in the distribution of objects, then the A_k relating to the various objects cancel, so, for example, it is possible that $\beta_k \to 0$ as $k \to 0$. More generally, assume

$$\beta(k) \sim k^{\alpha_*} , \qquad \alpha_* > 1 . \tag{13.3.4}$$

Substituting such a $\beta(k)$ into the formula for μ, we obtain

$$\mu \sim r^{-2} \sim \langle M \rangle^{-2/3} ; \qquad \Delta N \sim N^{1/3} .$$

The result is independent of the value of α_*. This is a disappointment: We had intended to measure μ and find out something about the spectral index α_*. The reason for the disappointment is contained in the definition of μ.

Taking a sphere with a given radius r, we consider fluctuations present in the case of the smoothest possible distribution of objects. Let the bulk density be constant over large scales. Even here though, the addition or removal of an isolated object lying near the surface of large sphere of radius R changes the total mass inside the sphere. The number of objects near the surface, N_s, is

$$N_s = 4\pi R^2 r_1 n , \tag{13.3.5}$$

where $r_1 \sim n^{-1/3}$ is the average distance between objects, and

$$\Delta N \sim \Delta N_* \sim N_*^{1/2} \sim N^{1/3} . \tag{13.3.6}$$

Small displacements of objects near the boundary can lead to strong fluctuations. Now, having a diagnosis, we can find an appropriate treatment.

In order to study the regularity of a distribution of objects on a given scale of order R, we shall attempt to determine

$$M(R) = 4\pi \int_0^\infty \rho \exp\left[-(r/R)^2\right] r^2 dr , \tag{13.3.7}$$

or, in the case of point masses,

$$M(R) = \sum_i m_i \exp\left[-(r_i/R)^2\right] . \tag{13.3.8}$$

It is very important that a sharp boundary be absent in the new definition. The effective volume is of order R^3, so formulae (13.3.7) and (13.3.8) correspond to $\langle M \rangle = \pi^{3/2} R^3 \langle \rho \rangle$. Note that now small displacements of the objects do not lead to a big change in $M(R)$.

With the new definition of M, let us carry out the calculation of $\langle M \rangle$ and of $\mu = (\langle M^2 \rangle - \langle M \rangle^2)^{1/2} / \langle M \rangle$ anew, averaging with respect to an arbitrary choice of the origin of the coordinates. Using the Fourier representation in which

$$\mu^2 = \frac{1}{2\pi^2} \int_0^\infty \beta^2(k) \exp\left(-k^2 R^2/4\right) k^2 dk , \tag{13.3.9}$$

we find

$$\mu \sim R^{-3/2-\alpha_*} \sim M^{-1/2-\alpha_*/3} , \tag{13.3.10}$$

to which corresponds $\Delta N \sim N^{1/2-\alpha_*/3}$. For $\alpha_* = 0$, i.e., for a flat spectrum, the "naive" result is obtained, a result that corresponds to a random, uncorrelated distribution of objects. Now, however, the effect of α_* on a reasonable definition of μ is apparent. For example, if $\alpha_* = 2$, then $\Delta N \sim N^{-1/6}$, so the absolute fluctuation (and not only the relative fluctuation $\Delta N/N$) decreases as $N \to \infty$.

On the other hand, a spectrum of perturbations which grows as $k \to 0$, $\alpha_* < 0$, is in principle possible (although it seems unlikely). We then require that

$$\int_0^\infty \beta^2 k^2 dk$$

converge at $k \to 0$. For this it is sufficient that $\alpha_* > -3/2$. Consequently, in terms of the index $\gamma \equiv \frac{1}{2} - \alpha_*/3$, one can state that the relative fluctuations $\Delta N/N \sim N^{\gamma-1}$ always decrease if the integral of the perturbation amplitude is finite. For a comparison with observations, especially when $\alpha_* > 0$, caution is necessary.

One must also be careful because the result is valid for the mass density

but not for the luminosity density L or for the number density of galaxies, the masses of which are not constant. Only the mass is a conserved quantity. Even if the mass is distributed homogeneously over large scales [which means that $\beta(k) \to 0$ for $k \to 0$], the ratio of the luminosity to the mass for separate galaxies, L_i/m_i, may be variable; this gives an extra contribution to the dispersion of luminosity in a given volume.

$$L(R) = \sum_i L_i \exp(-r_i^2/R^2) , \qquad (13.3.11)$$

Because of this contribution, $(\Delta L)^2 \sim L \sim R^3$ even if $(\Delta M)^2$ is proportional to M^γ with $\gamma < 1$.

Therefore, the use of Gaussian functions serve more as an aid to conceptual understanding than as practical tools since observations must deal with the variability among sources. We know of no practical methods for directly measuring the mass density, which is the quantity directly related to the initial perturbations.

13.4 The Limitations and Complexity of the Linear Theory

It is tempting to draw conclusions concerning the future and concerning incipient astronomical objects by assuming a given perturbation spectrum. Indeed, the comparative ease with which the evolution of the spectrum $\beta(k, t)$ is calculated makes it that much more tempting. So we must give some consideration to the difficulties and limitations that arise in such an approach.

Imagine that a density distribution is specified. For simplicity, imagine also that it grows, but remains similar to the initial distribution. One would like to take the region with $\delta > 0$ (or $\delta > \delta_0$, for some definite δ_0) as the precursor of a distinct object. One can then compute the number of such regions, the volume of each region, and the mass contained in each, but the final result must possess a statistical character. In other words, the final result must be given in the form of a distribution function of the objects with respect to mass and to such other properties, as the value of the velocity, the angular momentum, and so forth.

The first practical difficulty is that it is not an easy problem to calculate the properties of a surface of given $\delta = \delta_0$ for a distribution function whose Fourier expansion is specified (see Doroshkevich 1970). Indeed, though no solution is known in finite, analytic form, attempts at numerical experiments have been undertaken. In the two-dimensional case, for example, Peebles (1969*b*) proposed an ingenious method for modeling the experiment optically.

But one can still foresee several conceptual difficulties. If δ_0 is greater than the mean square deviation $(\langle \delta^2 \rangle)^{1/2}$, then the region with $\delta > \delta_0$ is small and is distributed in the form of small islands surrounded by a sea

with $\delta > \delta_0$. If, on the other hand, δ_0 is small or zero, then the topology of the parts with $\delta > \delta_0$ is not at all similar to the topology of known astronomical objects (e.g., clusters of galaxies). The surface $\delta = 0$ can be closed around an "object" with $\delta > 0$. It is equally probable, however, that a surface $\delta = 0$ can be closed around a "hole" surrounded by denser matter.

The root of the difficulties is that astronomical objects result from strong nonlinearities. These objects are indeed characterized by $\rho/\langle\rho\rangle \gg 1$; for example, $\rho/\langle\rho\rangle \approx 100$ for the least dense objects—clusters of galaxies. This completely destroys the symmetry between regions with $\delta > 0$ and those with $\delta < 0$, a symmetry that is characteristic of the linear theory. Evidently, a space with $\rho/\rho_0 > 50$, or $\delta = \rho/\rho_0 - 1 > 49$, occupies a fraction not larger than 2% of the space. But in no part of the space is $\Delta = -49$ (or even $\delta < -1$), because the density is everywhere positive.

We reiterate that the formation of astronomical objects must be described by nonlinear theory. Several hypotheses and results obtained using nonlinear theory are therefore set forth in the following chapter.

In order to finish with the statistical theory, it is necessary to mention some questions connected with the Fourier expansion of random functions in an infinite space. In a finite volume V, we worked with summations. The number of terms in the wave-vector interval $[k_x, k_y, k_z; k_x + dk_x, k_y + dk_y, k_z + dk_z]$ is $V dk_x dk_y dk_z/(2\pi)^3$. Now, when V increases, the number of terms in the expression

$$\delta = V^{-1/2} \sum_k A_k e^{i\mathbf{k}\cdot\mathbf{x}}$$

increases. But these terms have random magnitudes. Consequently, their sum is proportional to the square root of their number. Each term in the sum must be proportional to $V^{-1/2}$. It is necessary to keep these properties in mind in order to understand the meaning of the integral $\int A_k e^{i\mathbf{k}\cdot\mathbf{x}} d^3x$.

In the linear theory the evolution law was obtained for discrete waves, i.e., for separate terms in the sum. When, however, a transition is made to an infinite volume and an integral, it is necessary to keep in mind that the sum of squares over a definite volume in k-space plays the role of the square of the amplitude of a given wave:

$$\frac{\langle |A_k|^2 \rangle}{V} \to \frac{\beta^2(k)}{(2\pi)^3} d^3k = \frac{\beta^2(k) k^3}{2\pi^2} d(\ln k) . \qquad (13.4.1)$$

In the course of the expansion, k decreases as $(1 + z)$. For a single wave, meanwhile, the evolution law for A_k is

$$A_k(z) \sim t^{1/3} \sim (1 + z)^{-1} . \qquad (13.4.2)$$

Here we are taking the simplest possible case—that of a plane wave in the MD model. For the amplitude of the Fourier integral we therefore find

$$\Delta \sim |A_k|^2 \sim (1 + z)^{-2} \sim \beta_k^2 (1 + z)^3 , \qquad (13.4.3a)$$

where

$$\beta_k \sim (1 + z)^{-5/2}, \tag{13.4.3b}$$

or, more precisely

$$\beta(k, z) = \beta\left[k\left(\frac{1 + z}{1 + z_1}\right); z_1\right]\left(\frac{1 + z}{1 + z_1}\right)^{-5/2}. \tag{13.4.4}$$

Note that k is the wave vector in physical (not comoving!) space. We see again the convenience of the dimensionless mass functions $\Delta(M)$ and $\mu(M)$ introduced above (see eqs. [13.2.10] and [13.3.1]). The constant value of the argument M corresponds to the constant volume of the comoving space. Note that M does not include the contribution from radiation. In the course of evolution, Δ and μ change according to the same law that holds for the dimensionless amplitude of separate waves.

11 THE NONLINEAR THEORY OF PERTURBATIONS AND THERMAL INSTABILITY

11.1 Perturbations in a Matter-Dominated Model; Problems Admitting Exact Solutions

Three ways exist to approach the nonlinear problem. One way consists of an exact solution of particular problems with special initial conditions. Spherically symmetric perturbations can serve as an example of such problems. The second way consists of the construction of an approximation method that extrapolates the linear solution to the general case. This method is very powerful, but it is always necessary to remember the limits of its accuracy. The advantage of this method is that it is applicable to the case of nonsymmetric perturbations. Finally, the third way consists of a qualitative explanation of the properties of the general exact solution. To this method in particular is related the question of the singular points of the exact solution and that of the asymptotic behavior near these singular points.

Of course, a combination of the three methods gives the most reliable results. Here, as in general, an approximation method is good if it coincides with the exact solution for at least some special case and if it has the correct asymptotic behavior near the singular points.

The case of spherically symmetric perturbations can be analyzed exactly because the effect of neighboring perturbations vanishes. The second method takes account of the tidal action of neighboring perturbations, but the price to be paid for this is that the method is an approximation. We shall nevertheless examine both methods. Note that we shall here only consider the problem in which pressure forces are neglected. This is in accordance with the title of this section.

In the simplest spherically symmetric case, we consider a sphere with perturbed density Ω' on a Friedmann background with Ω. The density of the sphere is assumed higher than the average density ($\Omega' > \Omega$), but the mass of the sphere is equal to the mass which would be present in the unperturbed solution. That is, the perturbation occupies a sphere of radius R which consists of a sphere of radius r and density ρ', and a hole in the shell between r and R. The region outside R is strictly unperturbed.

If $\Omega \leq 1$, then the two cases $\Omega' \leq 1$ and $\Omega' > 1$ are possible. In the first case $\Omega' \leq 1$, the perturbed sphere expands indefinitely and a bound object

does not form. In the second, more interesting case $\Omega' > 1$, the perturbation behaves as a closed Friedmann model. Consequently, at some time t_m, the density in the perturbed sphere [of radius $r(t_m)$] reaches a minimum, the expansion ceases, and contraction begins. Formally, this contraction would lead to recollapse (infinite density) at the time $2t_m$.

The model of a Universe with a specific mean density Ω, but with holes of finite radius R containing pointlike masses in their centers (after the recollapse of the high-density region inside) was called a "Swiss cheese model" by Einstein and Strauss (1945). They stressed the point that in this special case strong local perturbations have no effect at all on the average expansion for which, therefore, an exact Friedmann solution applies. The paradox of infinite density disappears, however, if the initial state is not strictly symmetric, or, even with spherical symmetry, if that state is not homogeneous, i.e., if the density depends on radius, $\rho = \rho(r)$. For evolution with $P = 0$, there is in this case too a time when the density is infinite. The distinction is that the condition $\rho = \infty$ does not occur simultaneously over the whole volume. There is, rather, only an instant when this condition holds at some surface or at a single particle of the matter. At that instant shock waves arise, and subsequently the matter compressed by a shock wave has nonzero pressure $P \neq 0$. Before the emergence of a shock wave, the entropy is assumed to equal zero and the equation of state is assumed to be $P(\rho, S = 0) = 0$. After the emergence of a shock wave, $S > 0$ and $P > 0$, so oscillations about the finite-density equilibrium situation occur. Because of dissipative processes, the kinetic energy of the oscillations is transformed into heat. The amplitude of the oscillations thus decreases, a part of the mass can be thrown off, and the limiting equation of state of the perturbed sphere becomes that of an equilibrium configuration (Zel'dovich and Kazhdan 1970).

The total energy "of a part of a closed Friedmann model" is negative. At the time of maximum expansion, the kinetic energy of this part is zero (we neglect the pressure since the matter is dust) so that the total energy at that time equals the potential energy:

$$E_{\text{tot}} = -\tfrac{3}{5}\frac{GM^2}{r_m}. \tag{14.1.1}$$

According to the virial theorem, equilibrium at the same value of the total energy requires that the thermal energy equal $-E_{\text{tot}}$ and the potential energy equals $2E_{\text{tot}}$. Consequently,

$$U_{\text{eq}} = -\tfrac{3}{5}\frac{GM^2}{r_{\text{eq}}} = 2E_{\text{tot}} = -2 \times \tfrac{3}{5}\frac{GM^2}{r_m}, \tag{14.1.2}$$

so $r_{\text{eq}} = r_m/2$ and $\rho_{\text{eq}} = 8\rho_m$ (Peebles 1967*a*).

This result is an order-of-magnitude estimate, since Peebles neglects the mass loss. In an actual equilibrium configuration the density is not constant

with respect to the radius, so r_{eq} and ρ_{eq} are merely effective quantities. For a more detailed analysis, a complete study of the contraction after the time of maximum expansion is necessary. This is not difficult to do by numerical methods, but the results depend strongly on the specific assumptions made concerning the violation of the exact symmetry and concerning the departure from homogeneity.

One can predict that if $\rho(r)$ has a maximum at the center, then a shock wave arises during the contraction, and a small part of the matter is thrown off with a positive energy. The negative energy of the remaining, gravitationally bound matter increases in absolute magnitude. The equilibrium density also increases (Zel'dovich and Kazhdan 1970).

TABLE 10

TIME-DEPENDENT VALUE OF THE DENSITY PARAMETER δ DESCRIBING SPHERICALLY SYMMETRIC PERTURBATIONS

Time†	Value of δ in the Linear Theory	Value of δ in the Nonlinear Theory
$t \ll a^{*-3/2}$	$a^* t^{2/3}$	$a^* t^{2/3}$
$t = a^{*-3/2}$	1	4.5
$t = 2a^{*-3/2}$	1.7	∞

† The quantity a^* is a constant with the dimensions of (time)$^{-2/3}$.

It is instructive to compare the results of the nonlinear theory with the naive results from the linear theory. A simple calculation (details omitted here) leads to the following mnemonic rule for the case of an $\Omega = 1$ unperturbed model. At the time t_m, the perturbed density is

$$\rho'_m = \frac{(6\pi)^2}{64}\rho = 5.5\rho\,; \qquad \delta_m = 4.5\,. \tag{14.1.3}$$

In the linear theory, which yields $\delta = \text{const.}\ t^{2/3}$ and which gives the correct asymptotic behavior for $t \ll t_m$, we get

$$\delta_{\text{lin}} = \tfrac{3}{20}(6\pi)^{2/3} = 1.07 \quad \text{when} \quad t = t_m\,. \tag{14.1.4}$$

In the $\Omega = 1$ model, consequently, small, spherically symmetric, growing perturbations with the asymptotic behavior $\delta = a^* t^{2/3}$ for $\delta \ll 1$ halt the expansion at the time $t_m = 1.1a^{*-3/2}$, when $\delta_m = 4.5$, and lead to collapse of the perturbed sphere (i.e., to $\delta = \infty$ when $t_c = 2t_m = 2.2a^{*-3/2}$). See Table 10 for a summary of these results.

For the flat, dust model with $\Omega_0 = 1$, the total energy of each unperturbed

(spherical) part equals zero. Consequently, each growing perturbation with $\delta > 0$ has negative energy, which leads to the cessation of the expansion and to eventual collapse. More generally, the case $\Omega_0 < 1$ differs from cases for which $\Omega_0 \geq 1$. For $\Omega_0 < 1$, the unperturbed energy is positive; as was shown above, there then exists a critical perturbation amplitude needed to cause collapse.

It is interesting to examine the early moments of the solution, a time when the unperturbed density is large and the perturbations small, and to answer the question of what is the critical perturbation amplitude that leads to the formation of gravitationally bound bodies.

The answer is that the critical perturbation amplitude is

$$\delta_{\text{crit}} = \tfrac{3}{5} \frac{1 - \Omega_0}{\Omega_0(1 + z)} . \tag{14.1.5}$$

A mnemonic rule giving this value is that the critical perturbation amplitude (i.e., the amplitude characterizing a spherical perturbation which leads to $\rho_{\text{pert}} \to \infty$ and $\delta \to \infty$ in the nonlinear theory) is determined by the condition $\delta = 3/2$ as $t \to \infty$ for a calculation performed in the realm of the linear theory. For $0 < \Omega_0 < 1$, the linear-theory amplitude δ_{lin} at $z = 0$, necessary in order that a maximum of expansion occur in the nonlinear theory, satisfies $1.5 > \delta_{\text{lin}} > 1.07$. This δ_{lin} is the value of δ at the time the expansion stops. At the instant of collapse of the perturbed sphere, $\delta_{\text{lin}} = 1.5$ for $\Omega_0 = 0$ and $\delta_{\text{lin}} = 1.7$ for $\Omega_0 = 1$. If the perturbation amplitude is known at early times, then it is easy to determine the fraction f of matter now condensed. For $\Omega_0 = 1$, the asymptotic value is evidently $f = 0.5$. For $\Omega_0 = 1$ or $\Omega_0 \approx 1$, however, the limit $t \to \infty$ is not very interesting; we are interested in past or present observable situations, not the future. The asymptotic value for $t \to \infty$ is meaningful only for $\Omega_0 \ll 1$ inasmuch as the fraction of the matter condensed at the present time ($z = 0$) and at $t \to \infty$ is essentially the same in this case. This is because perturbations do not grow in the linear theory during most of the evolution for $\Omega_0 \ll 1$.

The general formula for the amount of condensed matter is complicated. A reasonable approximation is:

$$f = \frac{1}{\delta_0 \sqrt{\pi}} \int_{\delta_{\text{crit}}}^{\infty} \exp\left(-\delta^2/\delta_0^2\right) d\delta , \tag{14.1.6}$$

with $\delta_{\text{crit}} = 1.5$, $\delta_{\text{crit}} = 1.07$, or $\delta_{\text{crit}} = 1.7$, the exact value depending on Ω_0 and on whether δ_{crit} is determined by a maximum of expansion or by a subsequent collapse. Meanwhile,

$$\delta_0 = \delta(z = 0) = \delta(z) \frac{\Psi(z = 0, \Omega_0)}{\Psi(z, \Omega_0)} .$$

The function $\Psi(z, \Omega_0)$ determining the growth of the perturbations is defined in chapter 3. This function of the two variables z and Ω_0 can also

be rewritten as a function of the single variable $y = \Omega_0(1 + z)/(1 - \Omega_0)$. The initial value $\delta(z)$ at some z must be taken from observations or guessed from theory.

14.2 Perturbations in a Matter-Dominated Model; an Approximate Analysis of a General Asymmetric Case (the "Pancake" Model)

It is known from observations that in the expanding Universe the amplitude of density perturbations achieves a magnitude of the order of unity at that stage when the linear size of inhomogeneities is much less than the horizon size ct and much less than the radius of curvature a. Thus, a linear theory of perturbation growth ceases to be valid at a stage when the use of Newtonian physics remains valid, i.e., when relativistic effects are insubstantial. In addition, there are cases where the influence of gas pressure on perturbation growth is negligible. This is especially so in the case of the adiabatic perturbations whose evolution during the prerecombination, radiation dominated epoch was discussed in chapter 11. Below we consider the postrecombination conditions of such perturbations. Our task is to investigate nonlinear perturbations. The first step is to write the solution for linear perturbations in a special form. The next step will be to prove that this form provides a good approximation for large perturbations.

The assumed approximate solution (Zel'dovich 1970*b*) consists of a solution expressed in terms of Lagrangian variables. The Eulerian coordinates r (or x, y, and z) of the particles are written as functions of their Lagrangian coordinates s as

$$r = a(t)s + b(t)x(s) . \tag{14.2.1}$$

The first term corresponds to unperturbed motion. Neglecting the second term, we obtain

$$u = \frac{dr}{dt} = s\frac{da}{dt} = r\frac{1}{a}\frac{da}{dt} . \tag{14.2.2}$$

This is, of course, the Hubble expansion law. The Lagrangian coordinates s are thus defined as the comoving coordinates of the unperturbed motion.

The second term in formula (14.2.1) describes perturbations. This formula is exact in the case of small growing perturbations; we shall use it for finite and even large density contrasts. The function of the two variables s and t is factorized, i.e., it can be written as the product of the two functions $b(t)$ and $x(s)$.

Indeed, it was shown in § 10.3 that in the linear approximation and when $P = 0$, a perturbation of any form grows, but its form remains unchanged:

$$\delta = \delta_0\left(\frac{r}{a}\right)\phi_1(t) ; \quad w = w_0\left(\frac{r}{a}\right)\phi_2(t) .$$

At the same time, the spatial scale changes according to the general cosmological expansion. Here we ought to make one stipulation. The density distribution can be anything, i.e., there is no limitation imposed on the function $\delta_0(r/a)$. For a given function δ_0, however, the distribution of the peculiar velocity w_0 is connected with δ_0 by the condition $\nabla \cdot w_0 \sim -\delta_0$. A second condition that the velocity satisfies is $\nabla \times w_0 = 0$. This follows because a vortex velocity decreases rapidly in the course of the cosmological expansion and because a growing solution does not contain a vortex component. One can now reformulate these conditions in such a way that w_0 is derivable from an arbitrarily chosen potential field: $w_0 = \nabla\phi$. Then, $\nabla \times w_0 = 0$. We assume that the density satisfies $\delta_0 = -\nabla \cdot w_0$.

In order to pass from a linear solution to the proposed approximate solution, two steps remain. We first introduce the Lagrangian variable s, i.e., we shall follow the motion of each particle. In the unperturbed solution, $s = r/a$; furthermore, as was already noted, the perturbations depend on just this variable. When the perturbations are no longer small, i.e., on leaving the linear theory, we still use the variable s, which is no longer the same as r/a. The second step is related to the simple connection between the peculiar velocity and the density that exists in the linear theory only; the exact relation is $\partial\delta/\partial t = -\text{div}\,[w(1+\delta)]$, where the time derivative is taken moving with the Hubble flow. In the construction of the (approximate) nonlinear theory, we select as an extrapolation the linear formula $w = w_0(s)\phi_2(t)$ for the velocity. The corresponding expression for displacement has the form (14.2.1). We then compute the peculiar velocity and the density exactly:

$$w = \frac{dr}{dt} - Hr = \frac{da}{dt}\,s + \frac{db}{dt}\,x(s) - \frac{1}{a}\frac{da}{dt}\,(as + bx)$$

$$= \left(\frac{a(db/dt) - b(da/dt)}{a}\right)x\,.$$

The Hubble parameter is here $H = a^{-1}(da/dt)$. A relation is thus established between the functions ϕ_2 and b. Since the spatial dependence of $w_0(s)$ and of $x(s)$ is the same, $x(s)$ is also a vortex-free function.

Why is this particular variant of the linear approximation useful for an extrapolation into the region of finite and large perturbations? First, we will provide a qualitative answer. In the absence of any forces the exact solution is of the form $r = \text{const.}\ s + tv(s)$. In this circumstance the particle trajectories intersect, and infinite density is attained. Clearly the perturbations are large near this singularity. But, in the general case, the singularity takes the form of a two-dimensional surface. Only for degenerate cases does the intersection occur along a line or at a point. Now we take into consideration the gravitational forces. Near the two-dimensional singular surface these forces are *finite* (in contrast to the case of collapse to a

line or a point where they diverge). Thus, the gravitational forces do not exert a drastic influence on the growth of the perturbation, and they do not seriously modify the general picture. For a rigorous proof of this result see Novikov (1975). This is why it is reasonable to seek a solution for large perturbations in a form which is valid for small perturbations and small gravitational forces (eq. [14.2.1]). The important practical consideration is that if the displacement is given by equation (14.2.1), it is easy to perform the computation of the other quantities such as the density and the velocity. After studying the properties of such a solution, we will estimate its accuracy. We note that the motion of pressureless matter was also studied by Stanyukovich (1971); a useful introduction is given in the Russian book by Zel'dovich and Myshkis (1973). For the motion of dust in GTR, see Grishchuk (1967*a*, *b*).

Based on these considerations, we adopt equation (14.2.1) as the basis for our analysis. In the linear approximation, the function $b(t)$ is well known. Density perturbations are proportional to the ratio $b(t)/a(t)$, and this ratio is well known. The approximation consists of using formula (14.2.1) for a description of the perturbed motion even in the nonlinear stage and during collapse when infinite density is achieved for at least part of the mass. The practical advantage of this approximation is its mathematical simplicity: for a given $\mathbf{r}(\mathbf{s})$, the density equation in Lagrangian coordinates is exactly soluble.

Proceeding [we use $a(t_1) = 1$], first notice that $dm = \rho d^3r = \rho_1 d^3s = \langle\rho\rangle a^3(t) d^3s$. Here, $\rho = \langle\rho\rangle a^3[\mathfrak{D}(\mathbf{r})/\mathfrak{D}(\mathbf{s})]^{-1}$, with

$$a^{-3}\frac{\mathfrak{D}(\mathbf{r})}{\mathfrak{D}(\mathbf{s})} = \begin{vmatrix} 1+\frac{b}{a}\frac{\partial x_1}{\partial s_1} & \frac{b}{a}\frac{\partial x_2}{\partial s_1} & \frac{b}{a}\frac{\partial x_3}{\partial s_1} \\ \frac{b}{a}\frac{\partial x_1}{\partial s_2} & 1+\frac{b}{a}\frac{\partial x_2}{\partial s_2} & \frac{b}{a}\frac{\partial x_3}{\partial s_2} \\ \frac{b}{a}\frac{\partial x_1}{\partial s_3} & \frac{b}{a}\frac{\partial x_2}{\partial s_3} & 1+\frac{b}{a}\frac{\partial x_3}{\partial s_3} \end{vmatrix} . \qquad (14.2.3)$$

The quantity $\mathfrak{D}(\mathbf{r})/\mathfrak{D}(\mathbf{s})$ is a Jacobian, and its exact value is written above in terms of a determinant. We choose an appropriate system of coordinates (changing from point to point, but independent of time for any given particle) so that $\partial x_1/\partial s_1 = -\alpha$, $\partial x_2/\partial s_2 = -\beta$, $\partial x_3/\partial s_3 = -\gamma$, and $\partial x_i/\partial s_j = 0$ for $i \neq j$. The density can be rewritten in the form

$$\rho = \langle\rho\rangle/[(1-\frac{b}{a}\alpha)(1-\frac{b}{a}\beta)(1-\frac{b}{a}\gamma)] . \qquad (14.2.4)$$

Here α, β, and γ are functions of $\mathbf{s}$ only, while b/a is a function of t only. The ratio b/a is a universal function of time (depending also on Ω_0) and is conveniently expressed in terms of the redshift z. On the other hand, α, β, and γ depend on the specific form $\mathbf{x}(\mathbf{s})$ of the initial perturbation. The

quantities α, β, and γ characterize the deformation along the three orthogonal, principal axes of the deformation tensor. In the general case, $\alpha \neq \beta \neq \gamma$ and $\alpha \neq \gamma$, so for definiteness we choose $\alpha > \beta > \gamma$.

The case $\alpha < 0$ (so $\gamma < \beta < \alpha < 0$) is possible, but if $\alpha > 0$ (with the sign of the two other quantities arbitrary), the quantity $\alpha b/a$ grows and can reach unity in the course of the evolution; it follows from equation (14.2.4) that $\rho \to \infty$ here.

The infinite density arises as a result of the one-dimensional contraction along the axis corresponding to α. A quite definite picture thus emerges from our approximation: when the perturbations become sufficiently large, flat pancakes of collapsed dust form in various places.

The assumed solution (14.2.1) is, of course, not exact. Remember that the equations of motion in Lagrangian form are

$$\frac{d^2\mathbf{r}}{dt^2} = -\nabla\phi\,; \qquad \nabla^2\phi = 4\pi G\rho\,. \tag{14.2.5}$$

The first equation is linear in the displacements and in the potential. The second equation for the gravitational potential ϕ, as a function of the Eulerian coordinate $\mathbf{r}$, is linear as long as ρ is a function of the same Eulerian coordinates. But the expression for ρ in terms of $\mathbf{r}$ is a nonlinear function of the displacement, as is seen from (14.2.4). Only in the case of small perturbations can one write

$$\phi = \phi_0(\mathbf{r}) = \frac{4\pi}{6} G\langle\rho\rangle r^2 + \phi' \tag{14.2.6}$$

with ϕ' a linear function of $\mathbf{r} - a\mathbf{s}$. Here $(4\pi/6)G\langle\rho\rangle r^2$ is the unperturbed potential in the Newtonian approximation. The decisive argument in favor of the assumed approximation is that the infinite volume density ρ in the pancake corresponds to a finite density per unit surface area. This is a consequence of the one-dimensional contraction—in contrast to what happens in the spherically symmetric case. But an infinitely thin disk with finite mass and finite surface density μ leads to a finite potential (and, of course, to a finite field or potential gradient). Thus the error introduced by our approximation is finite even at the end of the contraction period, when $\rho \to \infty$. On the other hand, this approximate theory is asymptotically exact at the beginning of the evolution when the perturbations are small, inasmuch as the formulae used are the generalization of the linear theory. Numerical calculations for random initial conditions demonstrate that the errors at the time of formation of the pancakes do not exceed 20–30% on the average (Doroshkevich, Ryaben'kii, and Shandarin 1973; Doroshkevich and Shandarin 1974). Firmly established, however, is that the one-dimensional character of the contraction eventually results in the intersection of trajectories and the emergence of pancakes from the high-density matter. Note finally that the formation of a pancake is accompanied by the emergence of a pair of shock waves in the matter surrounding the pancake.

The physical applications of this theory are discussed in the following chapter.

It is necessary to emphasize once again that the picture set forth here is valid only when the pressure vanishes (or is negligibly small everywhere except at the pancakes). This proviso is, of course, suggested by the title of the section. An equivalent formulation is the following. The theory is applicable only to long wave ($\lambda > \lambda_J$) perturbations. The function s, which plays a major role in the approximate theory, is determined by the initial values of $\delta(s, t)$ and $w(s, t)$. These quantities must be separated into a growing mode, a decaying mode, and a vortex mode (the last satisfying $\nabla \cdot w = 0$ and $\delta = 0$). For $t \gg t_1$, however, only the growing mode survives. The case of initial vortex perturbations will be considered in § 15.9. The solution is described by formulae, in which, to be sure, x describes the amplitude of a growing mode. In this solution the motion is irrotational in r-space as well as in s-space, so that $x = \nabla_s \phi(s)$ and, therefore, the deformation tensor $\partial x_i / \partial s_k$ is symmetric.

One can show that, for perturbations dependent on only one Cartesian coordinate, s_1, for example, the solution is exact. The solution in this case takes the form

$$x = a(t)s_1 + b(t)x_1(s_1) ,$$

$$y = a(t)s_2 ,$$

$$z = a(t)s_3 .$$

The absence of vorticity imposes no constraints on the function $x_1(s_1)$. For vanishing pressure, such a solution is valid for any function $x_1(s_1)$ and for functions $a(t)$ and $b(t)$ satisfying the equations

$$\frac{d^2a}{dt^2} = -\frac{H_0^2\Omega_0 a_0^3}{2a^2} , \quad \frac{d^2b}{dt^2} = \frac{H_0^2 a_0^3 \Omega_0 b}{a^3} = -\frac{2}{a}\left(\frac{d^2a}{dt^2}\right) b ,$$

equations already well known to us. The solution for a describes the Friedmann expansion, while the solution for b [dependent on Ω_0 through the function $a(t)$ determined by the first equation] describes the linear state of any form of perturbations. In addition, however, this solution remains exact for large, one-dimensional perturbations until such time as particle collisions and the appearance of shock waves no longer allow one to neglect pressure.

Returning to the general three-dimensional case, for small values of b/a, one can write a first approximation as

$$\begin{aligned} \rho &= \langle\rho\rangle[1 + (\alpha + \beta + \gamma)b/a] ; \\ \delta &= f(s)b/a , \quad f = \alpha + \beta + \gamma . \end{aligned} \tag{14.2.7}$$

In this approximation a density perturbation grows but preserves its form; i.e., the character of the dependence of such a perturbation on the Lagran-

gian coordinate is given by a single function $f(s) = \alpha + \beta + \gamma$. For large perturbations, however, this similarity is destroyed. If $\delta < 0$, then, in accordance with the exact formula, $\rho > 0$ always holds, i.e., $-\delta < 1$. In agreement with the exact formula, in the case $\delta > 0$ an infinite density appears after a finite time, and not asymptotically as $b/a \to \infty$. At all points (more precisely, at particles) characterized by a given value of the sum $\alpha + \beta + \gamma$, the density changes in the same way in the early (linear) stage, but in different ways (with respect to the dependence on the values of the separate quantities, α, β, and γ) in the nonlinear stage.

Evidently, one can say that the proposed approximate theory takes into account not only the inflow of matter into the region where there is already a mass surplus, but also of the tidal action of neighboring masses, tidal action which indirectly affects the character of the contraction.

The rather subtle question of the probability of the appearance of a given set of coefficients α, β, and γ has been studied by Doroshkevich (1970). In a fixed coordinate system, the individual components of the tensor $\mathfrak{D}_{ik} = \partial x_i/\partial s_k$ consist of Gaussian distribution functions, if the distribution function of the initial perturbations is Gaussian and if the Fourier transforms of the perturbations are also Gaussian distributed. But the diagonalization of the tensor $\mathfrak{D}_{ik}$, i.e., the calculation of the three quantities α, β, and γ from the six quantities $\partial x_i/\partial s_k$, is a nonlinear operation. Consequently, the distribution functions of α, β, and γ are not Gaussian even if the distribution function of the $\partial x_i/\partial s_k$ is Gaussian.

Doroshkevich obtained as a distribution function the expression

$$P(\alpha, \beta, \gamma) = \frac{27 \exp\left(-\frac{3}{5}s_1^2 + \frac{3}{2}s_2\right)}{(8\sqrt{5})\pi\sigma_0^6}(\alpha - \beta)(\alpha - \gamma)(\beta - \gamma)\,; \qquad (14.2.8)$$

$$\sigma_0 s_1 = \alpha + \beta + \gamma\,, \quad \sigma_0^2 = \alpha\beta + \alpha\gamma + \beta\gamma\,.$$

Here σ_0 is the dispersion of the diagonal components of the tensor $\mathfrak{D}_{ik}$. The quantity σ_0 is expressed in terms of the Fourier components of the growing perturbations. Thus, σ_0 is proportional to $[\int M^{-2/3}\Delta(M, t)d(\ln M)]^{1/2}$. The factors $(\alpha - \beta)$, $(\alpha - \gamma)$, and $(\beta - \gamma)$ are of particular interest. The coincidence of α and β (or even of α and β and γ) representing transformation of a pancake into a cigar (or even into a point) is very improbable. Indeed, because of the presence of the factors indicated above, it is much less probable than the coincidence of two or three independent quantities. This fact further confirms the validity of the pancake picture.

It is necessary to note that contraction with respect to one coordinate can be accompanied by contraction as well as expansion in the plane of the pancake. As a matter of fact, only about 8% of the matter contracts along all three axes (which requires $\alpha > \beta > \gamma > 0$). However, 84% of the matter contracts in one direction but expands in one or two directions ($\alpha > \beta > 0 > \gamma$ or $\alpha > 0 > \beta > \gamma$). The value 8% is close to Oort's

estimate, which is that only about 1–6% of the matter is compressed in all three directions. It is, however, not evident that only this 8% eventually becomes gravitationally bound.

The fate of the pancakes is discussed later in chapter 15.

14.3 Nonlinear Spectral Theory

The case of dust is extraordinarily simple. In linear theory, the general, exact solution can be written straightforwardly without the use of Fourier transformations. It is moreover not surprising that in the nonlinear theory the approximate solution can be written in terms of functions dependent on spatial coordinates, multiplied by functions of time.

Now we shall consider the other case, that in which pressure is substantial. This case is particularly important for the analysis of processes occurring before recombination or small-scale entropy perturbations after recombination. It is necessary here to discuss the evolution of perturbations that propagate as sound waves or that lie near the boundary of the instability region. The spectral approach becomes inevitable under these circumstances. Thus, having begun with small-perturbation theory and having then constructed a linear theory, we must include additional effects from the beginning as corrections to the linear theory.

In the linear approximation, the equations for the Fourier amplitudes can be brought to the form (see, for example, [10.2.6]).

$$a\frac{d^2A_k}{dt^2} + b\frac{dA_k}{dt} + cA_k = 0\,, \tag{14.3.1}$$

where a, b, and c are functions of the wavelength. The homogeneity of equation (14.3.1) reflects the independence of the changes of every A_k from the other A_n—in the linear theory—and the possibility that an identically vanishing A_k is a solution.

The nonlinear corrections transform equation (14.3.1) into

$$a\frac{d^2A_k}{dt^2} + b\frac{dA_k}{dt} + cA_k = F(A_{k'}, A_{k''})\,. \tag{14.3.2}$$

It is the nonlinearity of the basic equations that gives rise to the term on the right-hand side. The most important result of taking account of the nonlinearity is the generation of harmonics absent in the initial spectrum. The conditions $A_k = 0$ and $dA_k/dt = 0$ at some instant of time do not guarantee that $A_k = 0$ at succeeding instants of time if the function F differs from zero.

Of course, an equation of the type (14.3.2) is an approximation. If one calls equation (14.3.1) the first approximation, then equation (14.3.2) can be called the second approximation. The exact equation, meanwhile, contains functions of the third order, such as $F(A_{k'}, A_{k''}, A_{k'''})$ and $F(A_k, A_{k'},$

$A_{k''}$), and also functions of all higher orders. Exact equations in the Fourier representation are such complicated integro-differential equations that they are useless. Exact equations in the ordinary spacetime form are more convenient since they are merely partial differential equations. We shall here limit ourselves to a discussion of the second approximation, equation (14.3.2).

The function F describes the interaction between two waves having wave vectors $\boldsymbol{k}'$ and $\boldsymbol{k}''$. This interaction produces a third wave having the wave vector $\boldsymbol{k}$, which is determined by the conservation law

$$\boldsymbol{k} = \boldsymbol{k}' + \boldsymbol{k}'' . \tag{14.3.3}$$

Consequently, the function F can be written in the form

$$F \approx \int d^3k' d^3k'' A_{k'} A_{k''} W(\boldsymbol{k}, \boldsymbol{k}', \boldsymbol{k}'') \delta(\boldsymbol{k} - \boldsymbol{k}' - \boldsymbol{k}'') . \tag{14.3.4}$$

Questions concerning nonlinear interactions have been considered in many articles. A good summary is contained in the monograph by Monin and Yaglom (1965, 1967) mentioned earlier. We now discuss the basic physical results.

1. If one neglects viscosity, then the interaction of longitudinal waves with one another does not lead to the emergence of transverse waves. This result follows from the exact theorem of the conservation of circulation. Transverse waves correspond to vortex motion without density perturbations, and are described by

$$\boldsymbol{W}_k = \boldsymbol{B}_k \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) , \quad \nabla \times \boldsymbol{W}_k = i\boldsymbol{k} \times \boldsymbol{B}_k \exp(i\boldsymbol{k}\cdot\boldsymbol{x}) . \tag{14.3.5}$$

Though $\nabla \cdot \boldsymbol{W}_k \sim \boldsymbol{k}\cdot\boldsymbol{B}_k = 0$ for transverse waves, $\boldsymbol{k} \times \boldsymbol{B}_k \neq 0$ here. But if $\nabla \times \boldsymbol{W}_k = 0$ initially, then this quantity remains zero subsequently if there is no viscosity and if shock waves do not arise. Consequently, in the quadratic (second-order) approximation, vortex-free motion only gives rise to vortex-free longitudinal waves.

2. Since the interaction of transverse waves with one another gives rise to longitudinal waves, density perturbations also arise in the second order. This result is also valid for higher orders. This effect lies at the foundation of vortex or turbulence theories of galaxy formation, which are discussed in more detail in the next chapter.

This effect also depends strongly on the equation of state. Thus, transverse waves do not depend in the same way on time in a plasma with $P = \epsilon/3$ as in dust with $P = 0$. More precisely, they decay because of the viscosity and the expansion. This detail, however, is not essential to the subsequent development.

In a medium with a large sound velocity b, short, longitudinal sound waves are periodic with the period $\tau = 2\pi/bk$. Transverse waves, meanwhile, do not lead to resonantly (with respect to time) growing longitudinal

waves, i.e., to an avalanche effect. Therefore, the amplitude of the short wave acoustic density perturbations generated is of the order of $\delta \sim u^2/b^2$, where u is the amplitude of the transverse motion.

But the effect grows sharply in dust and/or for any wavelengths, a fact used by Ozernoy and Chernin (1967, 1968). They assumed the existence of vortex (turbulent) motions in the precombination, RD plasma. After recombination, when the radiation pressure ceases to act on the matter, this turbulence leads to the emergence of large density perturbations.

3. The interaction of short-wavelength ($\lambda < \lambda_0$) longitudinal perturbations leads to the emergence of long-wavelength longitudinal perturbations (including growing density perturbations). If the initial spectrum falls sufficiently steeply on the long-wave side, then this process is the most important source of density fluctuations on large scales.

The exact result depends on the situation considered: whether one is studying perturbations whose wavelength is greater than the Jeans wavelength or is studying acoustic perturbations, traveling or standing sound waves. Under reasonable assumptions, one finds that the amplitude of the newly created (by the nonlinearity) long waves satisfies $A_k \sim k^2$, which leads to the law

$$\frac{\Delta M}{M} \sim (A_k{}^2 k^3)^{1/2} \sim k^{7/2} \sim M^{-7/6}. \qquad (14.3.6)$$

A simple explanation of this result is given in the following section, where a gas whose molecules are stars or clusters of stars is examined. Evidently, this is the extreme case of a huge number of short-wave perturbations, combined with the absence, initially, of long wavelength perturbations.

4. Longitudinal waves with a wavelength greater than some λ_0 generate short waves because of the nonlinear interaction, a process also important for cosmology. In accordance with the linear theory, viscosity suppresses long-wave perturbations corresponding to masses $M < M_J \approx 10^{12} M_\odot$. However, continuous generation maintains the perturbations in this wavelength interval.

Peebles (1970) has studied the process of shock-wave formation in a RD plasma, a process caused by just this generation of short waves. Thus, when a sinusoidal wave with no harmonics is transformed into a sawtoothed wave, a reasonable representation of a shock wave, the Fourier spectrum must contain many overtones. In particular, a discontinuity in the function $\rho(x)$, i.e., the inequality $\rho(x = x_s{}^+) \neq \rho(x = x_s{}^-)$, corresponds to a spectrum with A_k asymptotically proportional to k^{-1} (for $k \gg k_0$) in the one-dimensional case.

The emergence of a shock wave is also accompanied by the growth of entropy. In such a situation, entropy fluctuations remain even after the decay of the longitudinal acoustic waves.

Note too that in dust, where $P = 0$, the pancake formation examined

above is also accompanied by the emergence of high harmonics in the spectrum of density perturbations.

The theory of nonlinear wave interaction is far from complete at the present time. Only a few results important for cosmology have been fully worked out. It is therefore better to consider our discussion as an introduction to an unsolved problem written for future theoreticians. In this vein, we shall note several properties of the general theory.

From statistical mechanics we know that in an equilibrium situation each degree of freedom of a system has the same energy θ. Here, the effective temperature in energy units is $\theta = kT$, where k is the Boltzmann constant. (We shall avoid using this notation in order not to confuse the Boltzmann constant k with the wave vector k.) This means that $A_k = \text{const.}\ \theta^{1/2}$, with the constant independent of $\boldsymbol{k}$ and t. Therefore $A_k \sim \theta^{1/2}$ must be a solution to the kinetic equation corresponding to $\partial A_k/\partial t = 0$ on the left-hand side.

The equilibrium laws do not depend on the specific interaction mechanism. The creation of each wave with wave vector $\boldsymbol{k}$ from ones with vectors $\boldsymbol{k}'$ and $\boldsymbol{k}''$, for example, is accompanied by a reverse process, the decay of a wave $\boldsymbol{k}$ into waves $\boldsymbol{k}'$ and $\boldsymbol{k}''$ in our example. When all of the A_k have the equilibrium values, then the direct reaction is exactly balanced by the reverse reaction. This is a very general and very important principle. In particular, one can apply it to a system of equations describing longitudinal and transverse waves.

The equilibrium solution is strictly realized only in the trivial case of total equilibrium, when there are no macroscopic motions and fluctuations. But in those cases when there are macroscopic motions, the corresponding value of θ is extraordinarily large. For example, water moving at a speed of 1 cm s^{-1} on a scale of order 1 cm corresponds to $\theta \sim 1$ erg, or a temperature $T \sim 10^{16}$ K. But in this case it is evidently impossible for situations to occur in which all of the degrees of freedom would be characterized by such a temperature. After all, the water would evaporate at a temperature of $\sim 10^3$ K.

The theory of turbulence and the theory of fluctuations, on the other hand, always concern themselves with nonequilibrium situations. The amplitudes in a small part of the phase volume (for values of $\boldsymbol{k}$ much smaller than the inverse of the interatomic distance) are tremendously higher than the amplitudes of thermal motion. Therefore, we are usually concerned with a stationary, but nonequilibrium, situation.

The difference between these two possibilities is most clearly illustrated by the example of diffusion, whose governing equation is

$$\frac{\partial n}{\partial t} = -\frac{\partial j}{\partial x} = \frac{\partial}{\partial x}\left(\mathfrak{D}\frac{\partial n}{\partial x}\right). \tag{14.3.7}$$

In equilibrium, n = const and $j = 0$. But a stationary solution exists with j = const and $n = n_0 + j/v$.

In problems with turbulence or fluctuations, there is ordinarily a flow of energy to the short-wave side, where the energy transforms into heat because of dissipation. In situations where $A_k \to 0$ as $k \to 0$, in a part of the phase space, a flow of energy in the reverse direction is also possible.

One must continue to keep in mind the following important circumstance. Equilibrium is characterized only by the average value $\langle A_k \rangle$. But this is quite appropriate because the phase of the A_k is random and the distribution of the $|A_k|$ (the probability for finding a certain value of $|A_k|$) is Gaussian.

In the equations for the nonlinear interaction written above, it is tempting to make use of the same assumptions concerning spectral distributions. But the assumptions are not always admissible there. And this is not a question of little importance or pedantic reasoning. Recall, concerning the "pancakes," that there arises a density distribution having a large amplitude for short waves. These waves, however, are not random. Rather, they are correlated with the long waves, and all of them together form the flat pancake! Without this correlation, the same short-wave spectrum would give a quite different picture for the density distribution. Consequently, in problems similar to the problem of the emergence of pancakes and in problems concerning the formation of shock waves, the spectral approach is worse than the direct approach, i.e., than the analysis of the problems in spatial coordinates.

Finally, note that in a discussion of nonlinear problems, the spectral approach cannot be applied further than the second order in perturbation theory because it becomes extremely cumbersome and difficult.

14.4 The Emergence of Long-Wavelength Perturbations from a Gas of Stars or Star Clusters

Let us consider the important question of the formation of large scale clusters from discrete objects. This is the so-called theory of clustering.

There are two possible causes of clustering. The first is the presence of small perturbations of large linear dimensions in the initial perturbation spectrum. These perturbations grow according to the laws of gravitational instability. In due time their nondimensional amplitude reaches unity, which means that matter (including discrete objects of smaller dimensions formed earlier) is now clustered on the large scale. The second possible cause of clustering is due to nonlinear effects, which give birth to long-wave perturbations out of the short-wave ones. Below we will delineate the condition under which the first or the second cause is more important. We use the results of Doroshkevich and Zel'dovich (1975).

Assume that the perturbation spectrum (or at least the part of interest)

has an amplitude ΔM decreasing with growing linear scale, i.e., with growing M. Assume further that the rate of decrease is not too fast; this vague notion "not too fast" is made quantitative below.

In this case the small units—discrete gravitationally bound bodies—are formed first. But they are distributed in space nonuniformly, due to the presence of weaker perturbations on larger scale.

The formation of larger units depends on the initial spectrum, i.e., on the amplitude of the corresponding fluctuations. In the first approximation it is independent of the situation on the smaller scale. The growth of $10^{15}\,M_\odot$ perturbations is the same for hydrogen gas consisting of H atoms whose mass is $M = 10^{-24}\,\mathrm{g} = 10^{-57}\,M_\odot$ and for a gas, whose elementary units—"molecules"—are for example spherical star clusters with mass of the order $M = 10^5\,M_\odot$. Assume the density perturbations (to be exact, the growing mode amplitude which depends on both density perturbations and divergence of peculiar velocity at recombination) at recombination are given by a power law with cutoff

$$\delta_{\mathrm{rec}} = \frac{\delta\rho}{\rho} = b_0 M^{-\nu}\,, \quad M \geq M_1\,; \quad \delta_{\mathrm{rec}} = 0\,, \quad M < M_1\,. \tag{14.4.1}$$

The law of growth, in the simplest case $\Omega_0 = 1$, is given by

$$\delta = \delta_{\mathrm{rec}}\left(\frac{1+z_{\mathrm{rec}}}{1+z}\right) = \delta_{\mathrm{rec}}(t/t_{\mathrm{rec}})^{2/3}\,. \tag{14.4.2}$$

The approximate condition of formation of discrete gravitationally bound bodies is formulated in terms of the extrapolated linear theory density contrast, $\delta(M, z) = \delta(M, t) = 1$. Therefore the first appearance of bodies with mass M_1 occurs at time t_1 with corresponding z_1 given by

$$\delta = b_0 M_1^{-\nu}\left(\frac{t_1}{t_{\mathrm{rec}}}\right)^{2/3} = b_0 M_1^{-\nu}\,\frac{z_{\mathrm{rec}}}{1+z_1} = 1\,, \tag{14.4.3}$$

which gives

$$t_1 = t_{\mathrm{rec}} b_0^{-3/2} M_1^{3\nu/2}\,, \quad z_1 + 1 = z_{\mathrm{rec}} b_0 M_1^{-\nu}\,. \tag{14.4.4}$$

The growth of the characteristic mass occurs via the unification of smaller masses M_1 into a discrete bound aggregate with mass M, but this does not change the law:

$$M(t) = M_1\left(\frac{t}{t_1}\right)^{2/3\nu}\cdot \quad M(z) = M_1\left(\frac{1+z_1}{1+z}\right)^{1/\nu}\,. \tag{14.4.5}$$

In the special case of a "white noise" random initial fluctuation spectrum, $\delta \sim M^{-1/2}$, $\nu = \frac{1}{2}$ as considered by Press and Schechter (1974):

$$M \sim t^{4/3} \sim (1+z)^{-2}\,. \tag{14.4.6}$$

Therefore the average number of smallest bodies ("molecules" with mass M_1), which are bound together at a given moment, is growing proportional to $t^{4/3} \sim a^2$, i.e., to the square of the scale factor.

But the "white noise" spectrum is only one particular example of a power spectrum. It has no theoretical foundations in cosmology; other values of the power ν are possible. It is easy to rewrite the result for arbitrary ν. Below it will be shown that the law of independent perturbation growth leading to (14.4.5) is valid only if the spectrum is not too steep, $\nu < 7/6$. If this condition is violated ($\nu > 7/6$), one cannot neglect the nonlinear effects.

Let us consider the limiting case of an upper cutoff of the spectrum $\delta \equiv 0$ at $M > M_2$. The short-wave perturbations with $k > k_2$ (k_2 corresponds to M_2) in the process of growth are producing long-wave perturbations ($k < k_2$) by nonlinear interaction.

The Fourier amplitude of these newly produced modes is $A_k \sim k^2$. The k^2 foundation of the law is given below. First we give the physical consequences of this law (Zel'dovich 1965*d*; Peebles 1974*b*).

The corresponding density perturbation in a volume containing a large average mass M, much greater than the unit M_1, is given by

$$\delta(M) = (A_k{}^2 k^3)^{1/2}\big|_{k=M^{-1/3}} \sim M^{-7/6} . \tag{14.4.7}$$

By dimensional considerations we obtain from $\delta(M_1) = 1$, at $t = t_1$, the coefficient in the above formula:

$$\delta(M, t_1) = (M/M_1)^{-7/6} . \tag{14.4.8}$$

The final conclusion for the growth with time of bound masses in the expanding Universe is

$$M = M_1(t/t_1)^{4/7} = M_1\left(\frac{1+z_1}{1+z}\right)^{6/7} . \tag{14.4.9}$$

This law must be called a "self-similar clustering law." It is the asymptotic law for every initial spectrum which is steep enough (with exponent $\nu > 7/6$), i.e., for the case of absent or very weak long-wave perturbations.

The $\nu = 7/6$, $M \sim t^{4/7}$ law corresponds to the clustering of great masses, causally dependent (due to nonlinear effects) on the early formation of small masses M_1.

Contrariwise, the Press-Schechter law (14.4.6) depends crucially on assumptions about the initial perturbation spectrum. The power law for $M(t)$ depends on the assumed power law behavior of $\delta(M)$.

These considerations are especially important in the analysis of entropy perturbations. In this case the natural assumption is that the smallest

mass units M_1 are of the order of the Jeans mass after recombination. This mass is of the order of 10^5–$10^6\,M_\odot$ as long as the hydrogen temperature is equal to the RR temperature.

Assume for masses $M = 10^6\,M_\odot$ that the earliest moment of formation is at $z = 1000$. If the spectrum $\delta(M)$ is steep enough, we apply (14.4.7) to the clustering. This would give $M = 10^6(10^3)^{6/7} = 5 \times 10^8\,M_\odot$ now, at $z = 0$. This result is unsatisfactory: we know that clusters and superclusters of galaxies exist with much higher masses.

The Press and Schechter (1974) law $M \sim (1 + z)^{-2}$ gives $M_{\max} \approx 10^{12}\,M_\odot$, which is rather nearer to the mass of giant galaxies or small clusters of galaxies. As explained above, in this case the clustering on large scale depends on the initial long-wave perturbation. The law $\delta(M) \sim M^{-1/2}$ has no deep foundation. As a matter of fact, Gott and Rees (1975), taking $\delta(M) \sim M^{-1/3}$, obtain much better agreement with observations. There are various other ideas leading to large scale perturbations: the explosions of protostars (Doroshkevich, Zel'dovich, and Novikov 1967*a*), adiabatic perturbations with characteristic scale due to dissipative damping (Silk 1968; Sunyaev and Zel'dovich 1972*b;* Doroshkevich and Shandarin 1974). These questions are considered in detail in the next chapter. We now proceed to the derivation of the self-similar spectrum $\delta \sim M^{-7/6}$, $A_k \sim k^2$.

Assume that space is divided into small cells with every cell containing one or more particles whose distribution inside the cell is subject to three rules: (1) the mass in every cell is the same, (2) the center of mass of the particles in the cell coincides with center of the cell, and (3) the net momentum of the particles is zero if Hubble expansion is neglected, or it is equal to the cell mass times the Hubble velocity of the cell center. Obviously, any interaction inside the cell (even very nonlinear) does not violate these rules if they are imposed initially.

Taking the center of a cell as the origin of coordinates, the three rules are written (V = cell volume)

$$\sum_i m_i = \langle\rho\rangle V\,; \quad \sum_i m_i \mathbf{r}_i = 0\,; \quad \sum m_i \mathbf{v}_i = 0\,, \qquad (14.4.10)$$

the summation being done for masses inside the cell at the initial moment $t = t_0$. Thereafter, neglecting the Hubble expansion and gravitational interaction,

$$\mathbf{r}_i(t) = \mathbf{r}_i(t_0) + (t - t_0)\cdot\mathbf{v}_i(t_0)\,. \qquad (14.4.11)$$

The calculation of the Fourier coefficient A_k involves the sum

$$A_k \sim \sum_i m_i \exp(-i\mathbf{k}\mathbf{r}_i)\,. \qquad (14.4.12)$$

Let us calculate the contribution $A_k^{(0)}$ of the cell located around the origin for the longwave Fourier component. The wave vector being small, we decompose

$$\exp(-i\mathbf{k}\cdot\mathbf{r}) \approx 1 - i(\mathbf{k}\cdot\mathbf{r}) - \tfrac{1}{2}(\mathbf{k}\cdot\mathbf{r})^2, \qquad (14.4.13)$$

$$\begin{aligned} A_k^{(0)} = {} & \sum m_i - i\mathbf{k}\cdot\sum m_i\mathbf{r}_i - i\mathbf{k}(t-t_0)\cdot\sum m_i\mathbf{v}_i \\ & - \tfrac{1}{2}\sum m_i(\mathbf{k}\cdot\mathbf{r}_i)^2 - (t-t_0)\sum m_i(\mathbf{k}\cdot\mathbf{r}_i)(\mathbf{k}\cdot\mathbf{v}_i) \qquad (14.4.14) \\ & - \tfrac{1}{2}(t-t_0)^2\sum m_i(\mathbf{k}\cdot\mathbf{v}_i)^2 + \ldots . \end{aligned}$$

The first term equals $\langle\rho\rangle$; it is the same as if the matter distribution were strictly uniform with $\rho \equiv \langle\rho\rangle$. Obviously the sum of such contributions from all cells with appropriate multipliers $e^{-i\mathbf{k}\cdot\mathbf{R}}$ ($\mathbf{R}$ the coordinate of the other cells centers) is zero for $k \neq 0$.

The second and third terms (proportional to the first power of $\mathbf{k}$) are equal to zero for every cell. The nonzero terms start with the fourth; they are proportional to k^2. This is the proof that $A_k \sim k^2$, $A_k^2 \sim k^4$.

The result is very general. The rules on which it is based follow from the conservation of mass and momentum by the local interaction which is the source of perturbations. Forget about gravitation and gravitational instability. The thermodynamical equilibrium in a gas corresponds to the flat spectrum $A_k^2 = \text{const}$. The macroscopic dissipation rate by diffusion or viscosity is proportional to k^2. In order to preserve the equilibrium fluctuation amplitude it is necessary that the random motion excites the fluctuations at a rate which is also proportional to k^2. In the case of gravitationally interacting "molecules" the random fluctuations give initial amplitudes which are further amplified by gravitational instability.

In the expanding Universe the combination of an initial random force proportional to k^2 and instability growth $A_k \sim t^{2/3} \sim (1+z)^{-1}$ leads to the above mentioned clustering law $M \sim t^{4/7} \sim (1+z)^{-6/7}$.

Press and Schechter consider two cases: (1) The random occupation of cells by small bodies; in this case the first rule $\Sigma m_i = \langle\rho\rangle V$ is violated and therefore $A_k \sim k_0$, $\delta \sim M^{-1/2}$ ensues, giving the clustering law $M \sim t^{4/3} \sim (1+z)^{-2}$. (2) In the second case every cell contains one small body, so that the first rule is fulfilled: $\Sigma m_i = \langle\rho\rangle V$; but the initial position of the body inside the cell is random so that the second rule, $\Sigma m_i \mathbf{r}_i = 0$ is violated. In this case, as is seen from the expression for $A_k^{(0)}$, the first nonvanishing term is $A_k \sim |k|$. This leads to $\delta(M) \sim M^{-5/6}$ and the clustering law $M \sim t^{4/5} \sim (1+z)^{-6/5}$. The numerical calculations of clustering under gravitational forces by Press and Schechter confirm rather well the laws deduced above. The coincidence is better in the first case, $M \sim t^{4/3}$, than

in the second case $M \sim t^{4/5}$, probably because of the fluctuations involved with the sharp boundary and limited overall number of particles. The importance of boundary effects in calculating the density fluctuations especially for steep spectra ($A_k \sim k^2$, for example) was already noted earlier.

The simple clustering theory outlined above operates with the average bound mass only. The distribution of masses around the average value was analyzed by Doroshkevich (1967).

14.5 Thermal Instability and the Separation of a Homogeneous Gas into Phases

In the presence of energy sources and nonequilibrium processes, there exists a mechanism that causes instability in a homogeneous gas and that leads to the separation of a homogeneous gas into dense, cold clouds and a low-density, hot gas between the clouds. This mechanism, the so-called thermal instability mechanism, is unrelated to the action of gravity. As a consequence, its effectiveness, as measured by the increment of perturbation growth it causes, decreases for long waves. It is therefore not surprising that the action of thermal instability is characterized by the existence of an optimal scale size. We emphasize right away that the primary separation of the cosmological plasma into discrete bodies is not related to thermal instability since this instability requires energy sources. Conversely, under galactic conditions, where the gas density is high and energy sources are present, thermal instability is quite an important phenomenon.

This phenomenon, although strictly speaking not applicable to the realm of relativistic astrophysics, is important in the formation of stars and in other similar situations examined in this chapter. A summary of the basic results is presented below.

The thermal balance of a gas subject to the action of external sources of heating is described by an equation of the form

$$C_p \frac{dT}{dt} = h - \rho\psi(T) . \tag{14.5.1}$$

In many cases, in the first approximation, the inflowing energy per unit mass h does not depend on the density and the temperature of the gas. The loss of energy due to radiation, meanwhile, depends on two-body collisions between atoms or between an atom and an electron. Consequently, the loss rate is proportional to ρ. The temperature dependence $\psi(T)$ is rather complicated. For low temperatures ($\sim$100 K) the main process is the excitation of low-lying atomic levels of carbon and of other rare elements. For high temperatures, radiation at the Lyman-α line of hydrogen predominates, whereas, for still higher temperatures, free-free radiation is the main source of energy loss.

For a gas with a given composition and for given external sources of heating, one can compute the steady state temperature with the help of the equation

$$h - \rho\psi(T) = 0 . \tag{14.5.2}$$

Solving this equation leads to

$$T_s = \phi(\rho) .$$

The pressure follows simply as

$$P_s = P_s(\rho) = \frac{RT_s}{\mu}\rho \sim \rho\phi(\rho) . \tag{14.5.3}$$

For the conditions prevailing in our Galaxy, a computation gives non-monotonic behavior for $P_s(\rho)$. This behavior is qualitatively similar to

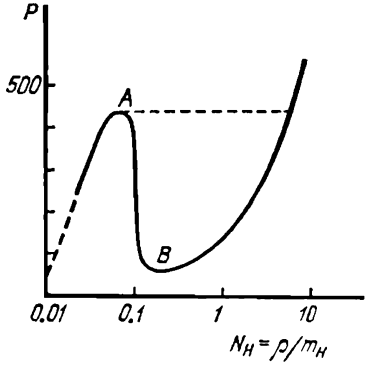


FIG. 45.—The dependence of the stationary pressure on density. The region labeled AB is unstable.

that of a van der Waals curve below the critical point (see Fig. 45). In the region between points A and B in Figure 45, where $dP/d\rho < 0$, the gas is evidently unstable! A formal calculation in the linear approximation confirms this. Thus, the perturbations grow as exp (ω_Tt), where, under these conditions,

$$\omega_T = k\left(-\frac{dP}{d\rho}\right)^{1/2} . \tag{14.5.4}$$

This formula is valid only so long as ω_T^{-1} is greater than the temperature relaxation time, a time which satisfies

$$t_T^{-1} \approx \rho\frac{d\psi}{dT} . \tag{14.5.5}$$

Such a situation bears the name "thermal instability" and is connected with the increase in energy loss which occurs when the pressure is constant, but the temperature falls and the density grows. The most important work in

this area was done by Field (1965), Goldsmith, Habing, and Field (1969), and Spitzer and Tomasko (1968).

The non-steady-state thermal-instability problem is also of interest. Consider a gas not in thermal equilibrium with the ambient radiation and in an environment devoid of heat sources. The gas cools, and one can ask how the perturbations change in the time required for cooling. If gravity is not taken into account, the short-wave perturbations grow only during a limited period of time. In the approach to complete equilibrium, therefore, stability necessarily sets in, and equilibrium temperature is established, and then the constant-pressure condition leads to a constant density. This, however, does not exclude the possibility of sharp growth in the perturbations at a stage far from equilibrium.

Let us return to the system with constant heating and in a stationary, nonequilibrium state. One can also study the final result of the perturbation growth in an exact way, not limiting oneself to the theory of small perturbations. This was done in Pikel'ner's (1967) important study. The result is that the system tends to a state in which the entire gas is in a stable configuration if the pressure is constant. This condition corresponds to a separation of the initially homogeneous gas into two phases L and N. Both phases are stable and exert equal pressure. The phase N is dense and cold while the phase L is dilute and hot. These phases form the basis of the theory that explains the origin of the clouds with a density $\rho \approx (3\text{–}6) \times 10^{-24}$ g cm^{-3} and a temperature $T \approx 70\text{–}100$ K. The clouds are surrounded by a hot gas with $\rho \approx 4 \times 10^{-26}$ g cm^{-3} and $T \approx (5\text{–}10) \times 10^3$ K. Estimates of the heating rate in the Galaxy give $h \approx 10^{40}$ ergs g^{-1} s^{-1} (Pikel'ner 1967). Notice that the solution is not unique. One can choose a horizontal line parallel to the line LMN, but lying lower or higher, to obtain another solution.

In the study done by Zel'dovich and Pikel'ner (1969), the question of the surface separating the dense, cold gas from the dilute, hot gas is examined with due regard for thermal conductivity. There then exists only one pair of points and one pressure for which this surface is stable. If one takes higher pressures, this leads to the condensation of the dilute gas into clouds. If, on the other hand, one takes lower pressures, this leads to the evaporation of the clouds. Consequently, there is only one completely stationary state.

Heat conduction is a slow process, and therefore a completely stationary state is not achieved in practice. The exact value of the pressure, the number, and the volume of the clouds formed in the manner just described can depend only on the details of the history of the system, and also on the presence of magnetic fields and of other complicating factors. The reader can find details in appropriate reviews.

Appendix to Section 14.2

Let us estimate the error in the approximate solution. This estimate can be given in an analytic form by making use of the following idea (Doroshkevich and Shandarin 1974). Using the approximate solution, we find the velocity and acceleration of each particle. Then, knowing the acceleration field, we find the corresponding gravitational potential. Finally, we find the density ρ_1 from this potential through Poisson's equation $\Delta\phi = 4\pi G\rho_1$. One can term this the "necessary" density in order to emphasize that this is the density distribution required to maintain the specified law for the particle motion. Specifying the density unambiguously specifies the potential for those conditions at infinity which follow from the absence of perturbations of infinite wavelength. Specifying the potential unambiguously gives the divergence of the accelerations, of the velocities, and of the displacements when account is taken of the initial conditions (for $t \to 0$, the displacements tend to zero). Finally, the displacements themselves are unambiguously determined by the divergence by virtue of the condition of the vanishing curl.

It is next necessary to compare the density ρ_1 with the density ρ calculated with respect to the displacements that follow from mass conservation. The coincidence $\rho_1 = \rho$ would signify that the solution is exact, i.e., self-consistent. One can therefore take the dimensionless ratio $\epsilon = (\rho_1 - \rho)/\rho$ as a measure of the error in the approximate solution. As will be shown below, ϵ grows in the course of time; it consists of two terms, one proportional to $t^{4/3}$ and one to t^2. At first, the density perturbations grow more slowly: $(\rho - \langle\rho\rangle)/\langle\rho\rangle \sim t^{2/3}$. In lowest order, $(\rho_1 - \langle\rho\rangle)$ and $(\rho - \langle\rho\rangle)$ coincide. The approximate theory coincides with the linear theory as long as the perturbations are small.

We now present the final results as given in the above cited works. First, recall that the solution is written in the form

$$r = a(t)q + b(t)\nabla\psi(q) .$$

The presentation of a perturbation in the form of the gradient of a scalar function guarantees the absence of a vortex component. Then

$$\langle\rho\rangle = \text{const.}\, a^{-3} :$$

$$\rho = \langle\rho\rangle \left| \delta_{ik} + \left(\frac{b}{a}\right) \psi_{ik} \right|^{-1}$$

$$= \langle\rho\rangle \left[1 + \left(\frac{b}{a}\right) J_1 + \left(\frac{b}{a}\right)^2 J_2 + \left(\frac{b}{a}\right)^3 J_3 \right]^{-1} ;$$

$$\epsilon = -3\left(\frac{b}{a}\right)^2 J_2 - 2\left(\frac{b}{a}\right)^3 J_3 .$$

Here $\psi_{,ik} \equiv \partial^2\psi/\partial q_i \partial q_k$, and the quantities J_1, J_2, and J_3 are the three invariants of the (symmetric) tensor composed of the second derivatives. It is most simple to write them after diagonalization of the tensor, when $\psi_{,11} = \alpha$, $\psi_{,22} = \beta$, $\psi_{33} = \gamma$, and the remaining components equal zero. Then, $J_1 = \alpha + \beta + \gamma$, $J_2 = \alpha\beta + \beta\gamma + \alpha\gamma$, and $J_3 = \alpha\beta\gamma$. But, precisely because J_1, J_2, and J_3 are invariants, it is easy to write them in an arbitrary, nondiagonal coordinate system as

$$J_1 = \psi_{,ii} ;$$

$$J_2 = \tfrac{1}{2}(\psi_{,ii})^2 - \tfrac{1}{2}\psi_{,ik}\psi_{,ik} = \tfrac{1}{2}J_1^2 - \tfrac{1}{2}\psi_{,ik}\psi_{,ik} .$$

$$J_3 = -\tfrac{1}{3}(J_1^3 - 3J_1J_2 - \psi_{,ik}\psi_{,kl}\psi_{,li}) .$$

The quantity b depends on time as prescribed by the linear theory of perturbations, $ad^2b/dt^2 + 2bd^2a/dt^2 = 0$. It is not essential that the model be flat; the formulae are valid also for the closed or open model, in which $(da/dt)^2 = c_1a^{-1} + c_2$, with $c_2 \neq 0$.

Notice now two important qualitative consequences. First, the relative error remains finite even when the density calculated according to the approximate theory is itself infinite. Both quantities, ρ_1 and ρ, go to infinity as $(1 - b|\alpha|/a)$ when $\alpha < 0$, i.e., according to the same law, but with somewhat different coefficients. Since the velocity and the displacement are obtained by integrating with respect to time, the error in them is several times smaller than the instantaneous value of ϵ.

Second, the fact noted above is confirmed: the solution given for the case of a perturbation depending on only one coordinate is exact. In this case, ψ depends only on q_1, so $\alpha \neq 0$ but $\beta = \gamma = 0$. Then $J_1 \neq 0$; $J_2 = J_3 = 0$.

Finally, let us prove that vorticity is absent in the motion described by the approximate solution. Inasmuch as the perturbation is described by a potential-type function $(\nabla\psi)$, the curl in q-space evidently equals zero:

$$(\nabla_q \times V)_3 = \frac{\partial V_1}{\partial q_2} - \frac{\partial V_2}{\partial q_1}$$

$$= \partial[(da/dt)q_1 + (db/dt)\partial\psi/\partial q_1]/\partial q_2$$

$$- \partial[(da/dt)q_1 + (db/dt)\partial\psi/\partial q_2]/\partial q_1 \equiv 0 .$$

However, it is necessary to prove that there is no curl

$$(\nabla_r \times V)_3 = \frac{\partial V_1}{\partial r_2} - \frac{\partial V_2}{\partial r_1}$$

in r-space either. In order to carry out the proof, it is convenient to choose coordinate axes at the particle studied in such a way that the tensor of

second derivatives of ψ is diagonal. In this system,

$$\psi = \psi_0 + n_1 q_1 + n_2 q_2 + n_3 q_3 + \frac{\alpha}{2} q_1^2 + \frac{\beta}{2} q_2^2 + \frac{\gamma}{2} q_3^2 + \ldots ;$$

$$r_1 = a q_1 + b \frac{\partial \psi}{\partial q_1} = (a + b\alpha) q_1 + b n_1 , \quad q_1 = (r_1 - b n_1)/(a + b\alpha) ;$$

$$V_1 = \frac{da}{dt} q_1 + \frac{db}{dt} \frac{\partial \psi}{\partial q_1} = \left(\frac{da}{dt} + \frac{db}{dt} \alpha \right) q_1 + \frac{db}{dt} n_1 .$$

The remaining components are expressed similarly. All of the non-diagonal derivatives are zero identically as long as no pancakes are formed, i.e., as long as $a + b\alpha \neq 0$. Then, for example,

$$\frac{\partial V_1}{\partial r_2} = \frac{\partial V_1}{\partial q_1} \frac{\partial q_1}{\partial r_2} + \frac{\partial V_1}{\partial q_2} \frac{\partial q_2}{\partial r_2} + \frac{\partial V_1}{\partial q_3} \frac{\partial q_3}{\partial r_2} = 0 ,$$

since

$$\frac{\partial q_1}{\partial r_2} = \frac{\partial V_1}{\partial q_2} = \frac{\partial V_1}{\partial q_3} = \frac{\partial q_3}{\partial r_2} = 0 .$$

The absence of curl in the approximate solution is thereby proved. This is an additional property of the approximate solution that coincides with a property of the exact, true, solution. The theoretical importance of this is evident in connection with the problem of galaxy rotation.

15 THEORIES OF GALAXY FORMATION

15.1 Introduction

The name of the chapter, referring not to one but to many theories, characterizes the state of work on the question. Indeed, within the realm of the theory of the hot Universe, several hypotheses are being worked out, are being compared with one another, and—we will not shrink from saying—are in competition with one another.

Common to these hypotheses is the assumption that in the RD period (before recombination) the plasma is almost homogeneous. The structure now seen in the Universe does not exist in that period. It is assumed, consequently, that the present-day structure arises after recombination ($z < 1400$). Now, in the phrase "almost homogeneous plasma," the emphasis belongs not on the word "homogeneous," but rather on the word "almost." What do deviations of the plasma state from strict homogeneity consist of? Recall the classification of the types of perturbations. Taking account of this classification, theorists have developed several different explanations of the deviations. We shall now briefly describe each such explanation.

A. Adiabatic Density Perturbations; Related Potential Motions and Metric Perturbations

The definition of adiabatic perturbations is that the ratio of numbers of baryons to numbers of photons, n_B/n_γ, is constant.

In the period when radiation is predominant, adiabatic density perturbations represent acoustic oscillations for $M \lesssim 10^{17} M_\odot$. After recombination, the perturbations grow in accordance with the law governing gravitational instability. An account of photon viscosity and of the damping of acoustic oscillations leads to a characteristic perturbation mass of the order of $10^{13} M_\odot$, which is close to the observed mass of clusters of galaxies. This coincidence is an argument in favor of the adiabatic theory. During the nonlinear stage of growth, disks (pancakes) arise and the gas is heated

by the shock waves it encounters. Sections 15.2–7 are devoted to a detailed analysis of the picture that emerges. Here notice only the great diversity of physical states of the gas in different parts of the pancake, a diversity offering possibilities for an explanation of such differing objects as quasars, elliptical galaxies, and spiral galaxies.

For a long time it seemed that the rotation of galaxies, which is manifested by their disklike or spiral structure, is impossible to explain in potential (vortex-free) theories of perturbations. Recently, however, Peebles (1967*a*, 1969*a*) showed that the angular momentum arising as a consequence of the gravitational interaction only is not far from observed values. Even more recently, Doroshkevich (1973) has taken into account modern ideas about the pancake stage in the process of the transformation of small perturbations into clusters of galaxies. Given this, the expected angular momentum is increased, due to the violation of the Kelvin-Helmholtz theory in shocks. What is important, vortex motion appears in part of the gas. In the presence of rotation, galactic magnetic fields then have a natural explanation.

One can also say that the theory of adiabatic perturbations uses rather small initial perturbations of the ideal cosmological model, perturbations which are necessary to obtain presently observed structure. Small—but nonzero (10^{-3} to 10^{-4} in dimensionless units)—metric perturbations are needed; they do not qualitatively change the local properties of the Universe in the early stages. The entropy is strictly homogeneous in this theory. Finally, it is the quasi-isotropic cosmological solution with a strictly constant matter composition that represents the state near the singularity leading to adiabatic perturbations.

B. The Entropy-Perturbation Model

The second variant of the theory, a variant discussed in § 15.8, assumes the existence of spatial fluctuations of the plasma composition. It is most natural to consider the fluctuations of the ratio of the number of baryons to the number of photons. Since the specific entropy per baryon is approximately equal to the number of photons per baryon, these fluctuations are also termed entropy fluctuations. In this conception, the metric and the dynamics of the expansion simply do not "feel" any perturbations at all during the entire stage of evolution when radiation is the predominant form. Baryon diffusion is small, so the baryon distribution specified initially remains unchanged up to recombination.

Immediately after recombination, the neutral gas at rest at a temperature of about 4000 K is not homogeneous. The temperature and average density at this time determine the Jeans mass $M_J \approx 5 \times 10^4\, \Omega_0^{-1/2}\, M_\odot$. This quantity is important inasmuch as large-scale perturbations grow according to a law that does not depend on scale if $M \gg M_J$. Smaller scale perturbations do not grow at all.

A falling initial mass spectrum for perturbations is most probable. In such a case, objects with the least mass consistent with the Jeans threshold form first. This minimum mass is somewhat larger than $5 \times 10^4\, \Omega_0^{-1/2}\, M_\odot$. In other words, gas pressure limits this mass from below. One can also assume that the role of pressure is substantial so that spherical bodies are formed, not pancake-shaped ones. Doroshkevich, Zel'dovich, and Novikov (1967*a*) have expressed the idea that this is how supermassive stars are formed. Such bodies evolve rapidly; the radiant energy and explosive energy given off evoke a chain of subsequent processes. In 1967 the present authors tried to obtain the entire structure of the Universe by proceeding from such entropy perturbations. Subsequent authors have posed the problem more realistically.

Dicke and Peebles (1968) suppose that objects with a mass of the order of 10^5–$10^6\, M_\odot$ transform into globular star clusters. Globular clusters consisting of stars with masses $\sim 1\, M_\odot$ are indeed observed as constituents of our Galaxy. Before the work done by Dicke and Peebles, the unusual similarity of such clusters—as manifested in the small scatter in their masses and densities—was unexplained.

But how does one explain the structure of the Universe on the largest scales? From the hypothesis of entropy perturbations? For if the initial amplitude of long-wave fluctuations (ones corresponding to masses $M \approx 10^{13}$–$10^{14}\, M_\odot$) is small enough, the subdivision of the gas into bodies with masses $M \approx 10^6\, M_\odot$ is not capable of leading to the observed large-scale structure. There is at least one way around this difficulty. By setting the initial spectrum of entropy perturbations "by hand" so that the amplitude at about $10^{13}\, M_\odot$ is just that required to obtain, at present, $\Delta M/M \approx 1$, one can obtain roughly the same conclusions as are obtained in the theory of adiabatic perturbations.

C. The Hypothesis of a Charge-Symmetric Universe

After the discovery of antiparticles and the creation of the physical theory of the charge symmetry of the properties of elementary particles, the idea that the whole Universe is charge-symmetric became very attractive. This attractiveness seized the widest strata of human society—from science-fiction writers and their readers to venerable scientists, inclusive.

This symmetry idea has been expressed in various ways so as to conform to various possible scales. The boldest statement is the assumption that both stars and antistars are present in the Galaxy and that pieces of antimatter (meteors and comets) have "invaded" the Solar System. See Rozhansky (1940*a*, *b*) and Konstantinov, Bredov, Belyavsky, and Sokholov (1966). The most cautious formulation says that certain clusters of galaxies consist of matter and others consist of antimatter.

In these charge-symmetric theories, one cannot avoid annihilation where the matter and antimatter intermingle. This results in the charac-

teristic reaction chain $p + p \rightarrow \pi^0 +$ (other particles), $\pi^0 \rightarrow 2\gamma$, with the gamma ray energy in the interval between 50 and 200 MeV. Specific sources of such gamma rays have not been found. However, a calculation of the expected background in various possible models of a charge-symmetric Universe is not simple. This is because one must specify still other assumptions in order to perform such a calculation.

An important argument against a charge symmetric theory of galaxy formation has been advanced by Bardeen (in an oral communication delivered by Field). In such a theory the regions occupied by matter and the regions occupied by antimatter are sharply separated. At a boundary the density of matter is zero, as is the density of antimatter. This means that the contrast in the matter density is always of order unity. Now the condensation of objects isolated from one another begins right after recombination. But in the situation under consideration here, the density of the objects formed must be of the order of the density of the matter at recombination, $\sim 3 \times 10^{-20}\,\Omega_0$ g cm^{-3}. Meanwhile, the density of the matter in galaxies is $\sim 10^{-24}$ g cm^{-3}, whereas the density of the matter in clusters of galaxies is of the order of 10^{-27} g cm^{-3}.

In work done by Klein and Alfvén (see the review by Alfvén and Elvius 1973), a charge-symmetric model of the entire Universe is examined. They assume the existence of a quite ingenious mechanism for the separation of particles and antiparticles by means of magnetic and gravitational fields. We emphasize, however, that the original state of the Alfvén model is a low-density plasma that is not in equilibrium with the radiation present. Alfvén thus renounces the hot cosmological model with a singularity, so that an explanation of RR in his theory is difficult.

Let us now turn from the conclusions to the origins of this type of theory. Modern physics takes as one of the great achievements of the twentieth century the discovery of charge symmetry, i.e., the symmetry between the properties (mass, decay lifetime, magnitude of electric charge) of particles and antiparticles. But the laws of physics are laws of motion, laws of evolution. At least for the present, a law governing the choice of initial conditions lies outside of physics. Therefore, the choice of an asymmetric Universe (with, for example, a baryon excess) is, *a priori*, no less compatible with the laws of physics than the choice of a symmetric Universe. Indeed, the discovery in 1964 of CP violation shows that such symmetry is not exact (for more details, see Part V).

Quite another charge symmetric theory has been investigated particularly intensively by Omnès (see § 6.3 and § 24.3). It is assumed, but not proved, that a very hot plasma with a large equilibrium density of baryon-antibaryon pairs (such as that assumed in the hot big bang theory) spontaneously decays into two phases, one with a baryon excess B and one with an antibaryon excess $\bar{B}$. The sizes of the initially separated regions are small; it is hoped that subsequently they will grow. The question is, can the ob-

served structure of the Universe be explained in this way? If this separation is real, it also must occur during the evolution of an asymmetric Universe.

One additional aspect of the question of charge symmetry is connected with the previous question concerning entropy fluctuations. Until we examine a plasma composed of photons and baryons, it is natural to make the assumption that somewhere baryons are more numerous and somewhere less numerous, though everywhere the density of baryons differs from zero and is of course positive. The photon density meanwhile is essentially the same everywhere because nonuniformities in the photon distribution are rapidly equalized.

When we turn to the period $t < 10^{-6}$ s, when $kT \geq m_p c^2$ and when there are many baryons and antibaryons (approximately as many as photons), $n_B/n_{\bar{B}} = 1.00000001$ on the average, and appearances change. It is necessary to adopt this initial ratio ($n_B - n_{\bar{B}} \approx 10^{-8} n_B$) because those baryons present today are the survivors of the annihilation which occurred when the temperature decreased below $\sim m_p c^2$. It is necessary to have an excess of baryons, but the excess must be small to obtain accord with present observations. Undoubtedly, the early stage of the hot Universe is quasi-charge-symmetric! But, there still must be a baryon excess and, moreover, in the theory of entropy fluctuations this excess is not spatially uniform!

In order for globular clusters (or supermassive stars) to form with a mass of the order of $10^6 M_\odot$, it is necessary, according to the remarks of § 14.2, that the baryon charge distribution have fluctuations on the order of more than 1% over the corresponding scale. Upon extrapolating back to the state existing at the time when there are many $B\bar{B}$ pairs, however, this only signifies fluctuations in the density ratio $n_B/n_{\bar{B}}$ within the limits 1.0000000099 and 1.0000000101.

In reality, we know the average ratio $n_B/n_{\bar{B}}$ itself (its deviation from unity is taken above to be equal to 10^{-8}) with very poor accuracy. It is possible that the ratio is several times smaller or larger. However, to order of magnitude, the fluctuation necessary for the emergence of clusters is correctly reflected in the numbers presented.

Notice immediately that these fluctuations are incredibly large in comparison with "statistical" fluctuations whose amplitude is of the order $N^{-1/2}$. Indeed, $M = 10^6 M_\odot$ corresponds to 10^{63} baryons, so $N^{-1/2} \approx 3 \times 10^{-32}$. But in the hot Universe with entropy we use $\Delta n_B/n_B \sim 10^{-10}$ before annihilation and $\Delta n_B/n_B \sim 10^{-2}$ after annihilation. Although these numbers are very different, both are much greater than 10^{-32}.

Although the fluctuations are certainly not "statistical," one can suppose that the smaller the scale, the larger the fluctuations. Then the assumption is not excluded that on some scale the fluctuations lead to an average baryon surplus of 10^{-8}. Limits on $n_B/n_{\bar{B}}$ given by $0.99999997 < n_B/n_{\bar{B}} < 1.00000005$ are then obtained. But this means that upon the completion of the annihilation process, there arise regions of matter and of antimatter

in which $n_B/n_\gamma \sim 5 \times 10^{-4}$ and $n_B/n_\gamma \sim 3 \times 10^{-8}$, respectively. Thus, the entropy fluctuation hypothesis (§ 15.8) is naturally linked with the hypothesis of charge symmetry. Indeed, the link involves the major prediction of the possible existence of antimatter regions. A further development requires consideration of the annihilation of these regions due to diffusive intermingling, consideration of the release of energy in this process, and consideration of the physical manifestations of this process.

These questions will all be examined not in connection with galaxy theory but rather later, in the chapter devoted to research on perturbations and their influence on RR.

Studies already undertaken have not turned up evidence of the annihilation process. Does there exist a formulation in which entropy fluctuations are automatically suppressed so that the absence of antimatter regions is ensured? One way to achieve that is to assume that at the singularity, the matter is cold and 100% charge-asymmetric, i.e., the matter consists of baryons only. Compare this with the recent work by Zel'dovich (1972*b*, 1973*b*) and with the material in § 24.9. We shall assume further that the enormous entropy presently observed is a consequence of some kind of physical process. Layzer and Hively (1973) expressed such a point of view even earlier. But the mechanism assumed by them seems questionable to us. Rees (1972) has examined the accumulation of entropy resulting from the decay of vortex motions. A new possibility is the formation and evaporation of primordial black holes (Hawking process). For more about this, see the following observation.

With the accumulation of entropy, antibaryons can also appear—but necessarily in the form of $B\bar{B}$ pairs. This guarantees the existence of an excess of baryons and the absence of regions with excess antimatter.

D. The Vortex (Turbulence) Theory

The origins of the vortex theory go back to the cosmogony of the solar system (the Kant-Laplace hypothesis). This rudimentary idea sought to explain the rotation of the planets and of the Sun itself, all in a single plane and in the same direction, by assuming that all of the bodies arise from the same rotating gas cloud.

Gamow (1952, 1954) assumed that the Universe is not only hot, but also turbulent. Von Weizsäcker (1951) developed the turbulence theory of the solar system and then a turbulence theory of the Universe too.

At the contemporary level, Ozernoy, Chernin, Chibisov, Nariai, Tomita, Silk, and others are developing the vortex theory with due regard (not subject to doubt) for the hot Universe.

It is assumed in the vortex theory that during the period when radiation is predominant, the plasma is in a state of turbulent vortex motion, a motion superposed on the general cosmological expansion. Inasmuch as photons are the main component of the plasma during this stage (not

only with respect to number density, but also with respect to energy or mass density), one speaks about "photon vortices," a concept explained in more detail by Ozernoy and Chernin (1967).

The turbulence theory will be examined in detail, with formulae and numbers presented in § 15.9 later. For the difficulty with this theory see § 15.10. Here we will give only its most general characteristics.

It is assumed that before recombination, the plasma (protons, electrons, and photons) is in motion. In the first approximation, the motion is considered incompressible. The motion is moreover presumed chaotic. In such a case, i.e., in the absence of any regulating motions except in the Hubble expansion, the motion must be vortex motion. It has recently been assumed that the velocity of this motion is significantly lower than the velocity of light and the velocity of sound in the plasma, and is of the order of $0.05c$–$0.1c$. The maximum scale of the motion is significantly less than that of the "horizon" at recombination.

The turbulence theory predicts a definite spectrum for the motion established. A definite condition connects the scale, the velocity, and the time required to establish the motion. The specification of one parameter—for definiteness, the maximum scale L over which a correlation exists—fixes the entire picture. The aforementioned condition then gives the magnitude of the velocity. The turbulence theory itself gives the velocity on scales smaller than L when one takes into account photon viscosity. On scales larger than L, a falling spectrum is assumed. It is also presumed that the turbulence is subsonic during the RD stage. This means that radiation pressure prevents a deviation of the density from the average.

We shall now give an intuitive estimate of the pressure and density fluctuations. Oppositely moving fluid volumes create a pressure difference of order $\Delta P = \rho u^2$, small in comparison with $P = \rho b^2$. Here b is the sound speed and u is the fluid speed. Correspondingly, $\Delta P/P \sim \Delta\rho/\rho \sim u^2/b^2 \ll 1$. A spreading out and breaking down of the flow thus occurs, so the motion hardly differs from the motion of an incompressible fluid. When recombination ensues, the pressure is shut off and the velocity of sound goes to zero. The motion then continues in accordance with its inertia, i.e., with a velocity distribution inherited from the previous stage. Meanwhile, the pressure is "incompressible" only at first. The incompressibility property is not preserved. As time passes, collisions between gas clouds disturb the inertial motion. The clouds first heat up because of compression and the effects of shock waves, but later they cool. Gravity then begins to act, and regions of increased density (a large portion, if not all, of such regions) transform into gravitationally bound objects.

The most important feature of the vortex theory, the rotation of the galaxies and of clusters of galaxies, is thus inherited from the initial vortex motion of the plasma before recombination. The presence of rotation does not require a special mechanism or a special explanation.

Those who have worked on the vortex theory have shown that the observed regularities pertaining to the density, the angular momentum, and the masses of galaxies and clusters of galaxies agree well with calculations. Recall too that only one numerical parameter enters the theory. Under these circumstances, the Kolmogorov theory describing the turbulence of an incompressible fluid (in the stage before recombination) and the hypothesis of supersonic turbulence (after recombination) play key roles.

What are the difficulties with the theory? For details see § 15.10. Let us begin with the past, i.e., with an examination of the plasma state long before recombination. During the RD stage when $P = \epsilon/3$, the speed of vortex motion does not change in the course of the expansion. During all of this time the velocity of the vortex motion can remain small in comparison with the velocity of sound $b = c/\sqrt{3}$ and the velocity of light c. Now it would seem that the ratio u/c is a measure of the deviation in the behavior from that of an unperturbed model. In reality, by taking account of GTR (see § 12.4), one sees that the metric perturbations depend not only on u/c, but also on the ratio of the scale size to the horizon size L/ct. It was shown explicitly that the vortex theory is incompatible with small perturbations to the Friedmann model (for more details, see Zel'dovich and Novikov 1970).

Is this not a difficulty with a deficiency in the vortex theory? Should we therefore dogmatically discard this theory? Today any answer to this question is surely subjective. The question remains up in the air because there is no existing detailed cosmological model corresponding to the vortex theory that traces the evolution of the turbulently moving plasma from the singularity up to recombination. It is not enough to say that the singularity is not "Friedmann-like." Those propounding the theory must give positive answers to the questions of what kind of singularity is necessary, with what properties, and with what implications for, for example, the ^{4}He formation, neutrino energy, and entropy production questions. This hypothetical, non-Friedmann-like solution must satisfy quite strict requirements. It must evolve as the expansion proceeds into a Friedmann solution with small perturbations and must evolve in such a way that the adiabatic perturbations are suppressed sharply in comparison with the vortex perturbations. Only by conforming to such a condition, whereby suppression of the adiabatic perturbations occurs in the period when the model is "Friedmannizing," is a period of photon vortices possible in the pre-recombination stage. Finally, the vortex theory also encounters specific difficulties concerning processes that take place after recombination (Peebles 1971*b*).

In short, according to an estimate made by Ozernoy, the essence of the matter is that large vortex velocities must lead to the formation of gravitationally bound objects during the very early stage when $z \sim 130$. The strong heat emission during this period leads to a high density of emerging

objects, for the supersonic turbulence damps rapidly and is not able to prevent the contraction. But the absence of noticeable RR perturbations (of its spectrum or of its isotropy) and the comparatively low density of the matter in clusters of galaxies, all force quantitative modifications in the vortex theory. It should be noted, however, that creators of the vortex theory explain the RR isotropy as due to scattering in the plasma after its second ionization. This secondary ionization occurs during the process of galaxy formation. For a recent calculation of this smoothing by scattering, see Ozernoy and Chernomordik (1976).

The direction of the modifications is such as to decrease the initial velocity of the vortex perturbations during the RD period. In the end, these quantitative changes change the picture of the phenomena qualitatively. Small vortex velocities, which would lead to the formation of objects late enough, in reality work in two stages. They produce at first small density perturbations which are then caught up in gravitational instabilities. In other words, because of gravity, potential motion develops, but in this case the vortex component can turn out to be insufficient to explain the observed rotation. Characteristic velocities of the plasma motion in the vortex theory significantly exceed the velocities in the adiabatic theory at recombination. In the first variants of the vortex theory, it was assumed that the formation of protoclusters occurs much earlier, when $z_c \sim 130$ (while in the adiabatic theory this is ordinarily assumed to occur when $z_c \sim 5$). But even under the assumption of equally small values of z_c, i.e., in the theory of weak turbulence, the initial velocities must be several times larger in the vortex theory than in the adiabatic theory. Indeed, it is difficult to reconcile the turbulence theory with RR observations with respect to this parameter.

An appraisal of the vortex theory does not ignore the state of other, competing viewpoints. Especially important is the solution to the question of the possibility of explaining the observed rotation of the galaxies by means of the theories of adiabatic and entropy perturbations. In those theories primordial vortex motions in the RD stage are not included.

We begin our review of the various theories by saying that the choice among them has still not been made. Refinements of the theories and discussions about them continue. The account above only illustrates this thesis and makes it concrete.

15.2 Adiabatic Perturbations: Prerequisites

The evolution of perturbations after recombination ($z \lesssim 1400$) is considered below. For a review see Doroshkevich, Sunyaev, and Zel'dovich (1973). The examination is carried to the stage of formation of gravitationally bound, isolated masses. These masses are presumably protoclusters of galaxies. The study thus includes the period when the perturbations

cannot be considered small and, therefore, when linear perturbation theory is not applicable.

Formally, one could specify for the density and the velocity of the neutral gas after recombination an arbitrary initial distribution and then study—analytically or numerically—the subsequent fate and evolution of the gas. In reality, however, it is important that one pose a narrower question and limit oneself to perturbations arising from initially adiabatic perturbations. Such a limitation permits one to develop an approximate theory for the process.

The properties of adiabatic perturbations used later are in essence these:

1. The motion is derivable from a potential, so the vorticity is zero initially and therefore remains zero later—up to the time a shock wave is formed.

2. The initial perturbations are small at recombination. After recombination a prolonged period follows during which the perturbations remain small and linear theory continues to be valid.

One can distinguish decaying from growing perturbations in the linear theory. In the general case their initial amplitudes are of the same order of magnitude, so that after a long period the decaying perturbations essentially disappear. The motion which continues into the nonlinear domain is that of the growing perturbation. Note, in particular, that the growing motion is vortex-free.

3. The linear theory for the evolution of perturbations is applicable to the period before recombination. It is natural to assume that at an early stage the amplitude of the perturbations is a more or less smooth function of wavelength.

The wavelength interval important for galaxy formation is not wide. In this interval, one can probably specify a power-law relation for the dependence of amplitude on wavelength. Thus

$$A_k \sim k^r ; \quad [\Delta(M)]^{1/2} \sim M^{-\nu} . \tag{15.2.1}$$

Here, $[\Delta(M)]^{1/2}$ is the quantity introduced earlier in § 13.2, so that the exponents r and ν are related by:

$$\begin{aligned} \Delta^{1/2} &= (A_k{}^2 k^3)^{1/2} \sim k^{(2r+3)/2} \sim M^{-1/2-r/3} ; \\ \nu &= \tfrac{1}{2} + \tfrac{1}{3} r . \end{aligned} \tag{15.2.2}$$

Damping related to the dissipative processes leads to a profound change in the form of the spectrum. One can write a smooth damping function in the form

$$\exp(-ak) \sim \exp(-R_0/R) \sim \exp[-(M_0/M)^{1/3}] , \tag{15.2.3}$$

where M_0 is of the order of $10^{13} M_\odot$. M_0 is strongly dependent on the Hubble parameter h_0 ($= H_0/100$) and on the density parameter Ω_0:

$M_0/10^{13}\,M_\odot = (\Omega_0 h_0^2)^{-2}\,[1 + 2.4\,(400\,h_0^2)^{-4/3}]^{-3/2}$. The exact form of the function is not very important. What this damping function means is that on scales corresponding to masses $M < M_0$ perturbations decay while those perturbations corresponding to masses $M > M_0$ do not.

4. After recombination a restructuring of the motion takes place. Growing perturbations develop that are characterized by the amplitude

$$[\Delta(M)]^{1/2} = M^{-n} \exp\,[-(M_0/M)^{1/3}]\,, \qquad n = \nu + \tfrac{1}{3}\,. \qquad (15.2.4)$$

The formulae presented here are valid for scales such that the perturbations become acoustic waves before recombination (i.e., for $M < M_2 \approx 10^{17}\,M_\odot$). The distinction between n and ν relates to the proportionality between $\delta\rho$ and ut/l for a growing mode in the postrecombination era. Here, u is the matter velocity in an acoustic wave before recombination, and $l \sim M^{1/3}$. Finally, recall that A_k is the Fourier amplitude of a density perturbation before decay. This amplitude coincides (to order of magnitude) numerically and in its dependence on the wave vector with the amplitude of a scalar metric perturbation. Moreover, A_k is time-independent for the entire period from the singularity until the instability sets in. It seems likely that $\nu = 0$, $n = 1/3$, and $r = 3/2$ (see the discussion below of the entropy theory). Other values for ν, n, and r are not, however, excluded.

5. After recombination one can neglect the pressure of the neutral gas when studying the evolution of perturbations. The basis for this is that $M_0 \gg M_J$, where M_J is the Jeans mass for the neutral gas (see § 15.3 below). Thanks to this, the approximate theory for the growth of perturbations in a dust medium, a theory set forth above in § 14.2, is applicable here.

Now let us turn to the motion following from the first and fifth points enumerated above. In the discussion of the approximate theory, it was explained that regions of infinite density arise in the course of time. After the emergence of such a region, the nearest gas particles encounter the dense region and are stopped. We will therefore examine in detail below the motion of the gas, its heating by shock waves, and the subsequent fate of the compressed gas.

15.3 Shock Waves

We shall here turn to the approximate solution. The quantity α (see formula [14.2.4]) determines the moment when the density reaches infinity (and, indeed, whether it does at all) through the condition

$$\alpha(q)\,\frac{b}{a} = 1\,. \qquad (15.3.1)$$

Consider a particle with a specific value q_m at which α is maximal: $\alpha = \alpha_m$. It is sufficient that this be a local maximum, i.e., that $\alpha(q_m)$ be greater

than the value of α at nearby points, even though there can exist other maxima, far away, some of which are greater than α_m.

Near q_m one can expand α as a function of q in a series:

$$\alpha = \alpha_m + A_{ik}(q_i - q_{m,i})(q_k - q_{m,k}) + \dots . \qquad (15.3.2)$$

The first-order terms $(q_i - q_{m,i})$ are absent because the quantity α is a maximum at the particle q_m. For the same reason, the quadratic term is everywhere negative. Thus the surface with constant α is an ellipsoid with the particle q_m at the center. It follows from dimensional considerations that $|A_{ik}| \sim \alpha_m/R_c^2$, where R_c is the characteristic size of the perturbations that avoid decaying in the period before recombination.

At the definite moment t_m, which is determined by the condition $\alpha_m b(t_m)/a(t_m) = 1$, an infinite density arises—at first at the particle q_m. Then the matter contracts with a smaller and smaller value of α. It is not difficult to convince oneself that all three axes of an ellipsoid in which the condition $\alpha(q)b(t)/a(t) = 1$ is achieved have lengths (in order of magnitude)

$$r \sim R_c\left(\frac{t - t_c}{t_c}\right)^{1/2} . \qquad (15.3.3)$$

The mass of the ellipsoid thus grows in proportion to $(t - t_c)^{3/2}$.

But the ellipsoid has three equal axes (in order of magnitude) only in q-space—in Lagrangian coordinates. In Eulerian coordinates—r-space—the ellipsoid is actually contracted in the direction in which the deformation takes place quickest of all. This is the direction along which the deformation rate is given by the quantity α_m (if $\alpha > \beta > \gamma$; see § 14.2).

The transition from Lagrangian to Eulerian coordinates is accompanied by a contraction of the ellipsoid. Upon a contraction of the ellipsoid in any direction one obtains a flat figure—an ellipse—or, in our parlance, a pancake, with $\rho = \infty$. At each element of the surface of the pancake the contraction takes place independent of the evolution at nearby elements. This is because all gradients in the direction normal to the plane of the pancake are much greater than the gradients in the plane of the pancake. One can therefore examine a whole collection of local questions—the laws for the changes in pressure, density, and temperature—in a one-dimensional approximation (see Zel'dovich 1970*b*; Sunyaev and Zel'dovich 1972*b*).

Consider then a one-dimensional model with a single sinusoidal wave and with $\Omega_0 = 1$. The general solution (see § 14.2) is exact in this case!

Let

$$r = a_0 t^{2/3} q - \frac{\kappa}{\kappa} b_0 t^{4/3} \sin(\kappa \cdot q) , \qquad (15.3.4)$$

i.e.,

$$\begin{aligned} x &= a_0 t^{2/3} q_x - b_0 t^{4/3} \sin(\kappa q_x) , \\ y &= a_0 t^{2/3} q_y , \quad z = a_0 t^{2/3} q_z . \end{aligned} \qquad (15.3.5)$$

The density is

$$\rho = \langle\rho\rangle\left[1 - \frac{\kappa b_0}{a_0} t^{2/3}\cos(\kappa q_x)\right]^{-1}; \qquad \langle\rho\rangle = (6\pi G t^2)^{-1}. \tag{15.3.6}$$

At first, therefore, the density goes to infinity in the plane $q_x = 0$, $x = 0$ at $t_c = [a_0/(\kappa b_0)]^{3/2}$. In the subsequent development, we consider a single cell of the periodic array. Thus $-\pi/\kappa \leq q_x \leq \pi/\kappa$ and also $(2n-1)\pi/\kappa \leq q_x \leq (2n+1)\pi/\kappa$. The corresponding value of z is called z_c. To simplify the notation, we write z everywhere instead of $z + 1$, although for $z \sim 3$–5 the reader can easily make the correction. This allows us to write $z/z_c = (t_c/t)^{2/3}$.

At $t > t_c$ let two shock waves arise and propagate symmetrically from the origin. Let us first examine the case in which the gas cools instantaneously. The density is then infinite, and the surface of the shock wave does not move away from $x = 0$. Now let us find the law which governs the degree of contraction of the matter and the pressure of the shock wave. The shock wave and the contracted matter are put at $x = 0$, from which it follows that

$$\frac{\sin \kappa q_x}{\kappa q_x} = \left(\frac{t_c}{t}\right)^{2/3} = \frac{z}{z_c}. \tag{15.3.7}$$

This equation has a nontrivial ($q_x \neq 0$) solution for $t > t_c$, or $z < z_c$. It is convenient to introduce the quantity $\mu = \kappa q_x/\pi$. This quantity, a type of Lagrangian coordinate, is the ratio of the mass included between the origin $q_x = 0$ and a given q_x to the total mass in half of a period cell. We then have

$$\frac{\sin \pi\mu}{\pi\mu} = \frac{z}{z_c}. \tag{15.3.8}$$

For $\mu \ll 1$, this gives

$$1 - \tfrac{1}{6}(\pi\mu)^2 = \frac{z}{z_c}, \qquad \mu = \frac{1}{\pi}\left[6\left(1 - \frac{z}{z_c}\right)\right]^{1/2} = \frac{2}{\pi}\left(\frac{t}{t_c} - 1\right)^{1/2}. \tag{15.3.9}$$

The law giving the dependence of the mass of the compressed gas on time, $\mu \sim (t - t_c)^{1/2}$, differs from the general law $m \sim (t - t_c)^{3/2}$ found above. This is because here a degenerate case is being considered. This case consists of the formation of a one-dimensional shock wave along the axis $x = 0$. In the general case, the amount of matter at each point of the pancake's surface grows as $(t - t'_c)^{1/2}$, with different values of t'_c at different points in the plane of the pancake [whose surface grows as $(t - t'_c)$].

Further, $\mu = \frac{1}{2}$ for $z = 2z_c/\pi$. According to the assumptions made, all of the matter passes through a shock ($\mu \to 1$), but only asymptotically, after an infinite time, when $z \to 0$. (To be exact, when $t \to \infty$, $z \to -1$.)

It is particularly necessary to notice what matter is subjected to the

shock wave induced compression. It is the matter which does not manage to achieve infinite density as the neighboring layers approach and also that matter which is expanding. For $t \ll t_c$ (during the linear stage) half of the matter contracts and half of the matter expands. In Table 11 is given the mass fraction μ compressed by the shock for $t > t_c$. Table 11 thus presents the characteristic dynamics of the compression under the assumptions made.

We shall now find the speed at which the shock wave strikes the matter and the matter density in front of the shock wave. Convenient formulae are obtained if one chooses μ for the independent variable; then one need not solve a transcendental equation. A simple computation then gives

$$
\begin{aligned}
|u| &= \frac{\lambda_c}{t_c}\frac{1}{3\pi}(\pi\mu \sin \pi\mu)^{1/2}; \\
\rho &= \rho_c \frac{\sin^3 \pi\mu \tan \pi\mu}{(\pi\mu)^3 (\tan \pi\mu - \pi\mu)}; \\
P &= \rho u^2 = \rho_c \left(\frac{\lambda_c}{t_c}\right)^2 \frac{\sin^4 \pi\mu \tan \pi\mu}{9\pi^2 (\tan \pi\mu - \pi\mu)(\pi\mu)^2}.
\end{aligned}
\tag{15.3.10}
$$

Here the notation $\lambda_c = 2\pi a_0 t_c^{2/3}/\kappa$ is introduced. This is the wavelength of a perturbation at the characteristic time t_c; $\rho_c = (6\pi G t_c^2)^{-1}$ is the average density at this time.

A characteristic feature of the formulae presented above is their behavior for small values of μ, i.e., for small values of $t - t_c$ or $z_c - z$. These values correspond to a time soon after the formation of the shock wave. We find

$$
\begin{aligned}
u &\sim \mu \sim (t - t_c)^{1/2} \sim (z_c - z)^{1/2}; \\
\rho &\sim \mu^{-2} \sim (t - t_c)^{-1} \sim (z_c - z)^{-1}; \\
P &= \text{const}.
\end{aligned}
\tag{15.3.11}
$$

The last formula does not mean that the pressure is strictly constant; its derivative is not equal to zero. But it is important that the pressure does not have a power-law dependence on μ. The pressure is neither zero nor

TABLE 11

CHARACTERISTICS OF SHOCK-WAVE COMPRESSION

t/t_c	z/z_c	μ
1.0	1.0	0.0
1.025	0.983	0.1
1.33	0.83	0.333
1.96	0.64	0.5
6.1	0.30	0.76
28	0.1	0.9
∞	0.0	1.0

infinity at the initial stage. In fact, with the passage of time, the pressure decreases, and

$$P(\mu = \tfrac{1}{2}) = \frac{4}{3\pi^2} P(\mu = 0) = 0.134P(\mu = 0) . \qquad (15.3.12)$$

Now let us consider the opposite limit, which constitutes another assumption about the thermal behavior of the gas—the complete absence of heat losses and transfer. A description of the motion after the appearance of a shock cannot now be constructed in an elementary way without a numerical computation, even given the simplifying assumptions made above. All that one can find simply is the asymptotic behavior at the beginning of the contraction, when $\mu \ll 1$. As was the case before, the pressure has a definite limiting value which is, however, larger than in the first case by the factor $11/9 = 1.22$. In the same limit, the amount of matter compressed by a hot shock (without heat losses) is $(11/9)^{1/2} = 1.1$ times larger than the amount of matter compressed by the shock with instantaneous cooling of the gas. This is because in the hot case the shock propagates from the plane $x = 0$ toward the infalling matter.

The average density of the compressed matter is 12 times larger than the density of the matter in front of the shock before compression (i.e., three times greater than immediately after the shock). Figure 46 shows the density distribution at t_c, i.e., at the time when $\rho = \infty$ is achieved in the plane $x = 0$. On the next figure shown is the density distribution at $t = 7t_c/6$, when about one-fourth of the matter is condensed. Figure 47 applies to the case of rapid cooling, while Figure 48 applies to the adiabatic case.

The characteristic feature of the adiabatic case has already been mentioned. This is just the breaking away of the shock from the plane of symmetry. At first (when $\mu \ll 1$) the path of the wave is short. As a matter of fact, the average density behind the shock front (i.e., the density of the compressed matter) is $\rho_{sh} = \langle\rho\rangle\, 33/(\mu\pi)^2$. Hence the thickness of a compressed layer divided by the size of a periodic cell (by the distance between nearby compressed layers) is

$$\alpha = \frac{2x}{\lambda_c} = \frac{\mu\langle\rho\rangle}{\rho_{sh}} = \frac{\mu^3\pi^2}{33} . \qquad (15.3.13)$$

From these formulae we find that when 10% of the matter is compressed by the shock ($\mu = 0.1$), this matter occupies a fraction $\alpha = 3 \times 10^{-4}$ of the volume. This means that its density is 300 times greater than the average density. If instead 30% of the matter is compressed, we obtain $\alpha = 10^{-2}$ and $\rho/\langle\rho\rangle = 30$. We emphasize that the calculation has been performed under the assumption of strict adiabatic behavior, hence without any account of heat loss. The true value of α (for a given type of perturbation) can only be smaller and the true density only larger than the quantities presented.

In the future course of the process, a mutual attraction among the layers of the compressed matter switches on, leading to the existence of a gravitational pressure. The pressure at the front of the shock arises because of the momentum of the infalling matter. Given the new circumstance, the pressure in the middle of the plane between two shocks (for $x = 0$) is greater than the pressure of the shock by the magnitude of the gravitational pressure, which is

$$P_g = \tfrac{1}{2}\pi G\sigma^2 . \qquad (15.3.14)$$

Here σ is the surface density (in g cm^{-2}) of the whole layer. This means the matter from $-\mu$ to $+\mu$ when account is taken of the matter to the left and to the right of the origin of coordinates. As the amount of compressed matter (the dimensionless quantity μ) increases, the ratio of the gravitational pressure to the shock pressure also increases. Numerical calculations performed by Doroshkevich and Shandarin (1973) show that when $t \sim 4t_c$ the process dies away. The compressed matter is held in place by gravity, the falling of matter is suspended, and the shock weakens. An unrealistically large time would be needed in order for the shock to com-

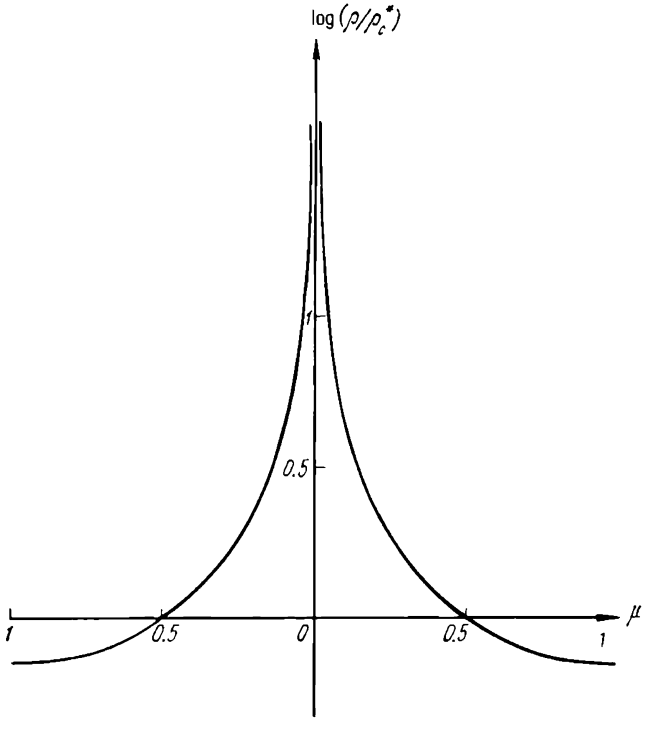


FIG. 46.—The density distribution (density ρ as a function of Lagrangian coordinate μ) in a developing cluster at the time $t = t_c$. At the location $\mu = 0$, the density is infinite; cf. eq. (15.3.6).

press all or even 90% of the matter. Meanwhile, the new processes of radiation, the birth of stars, and so forth are beginning in the Universe. These, of course, are not accounted for in the schematic picture of a one-dimensional adiabatic compression. Therefore, it is advisable here to call a halt to the hydrodynamical calculations.

15.4 Thermal Processes in the Compressed Gas

In the adiabatic approximation, the temperature of the gas compressed by the shock front is given by the expression

$$T = 2.5 \times 10^6 z_c \mu^2 \left(\frac{M}{10^{13}\, M_\odot}\right)^{2/3} \qquad (15.4.1)$$

for asymptotically small values of μ. The numerical value presented is for a sinusoidal wave, for an identically zero initial temperature, and in the complete absence of heat losses and transfer. The temperature first

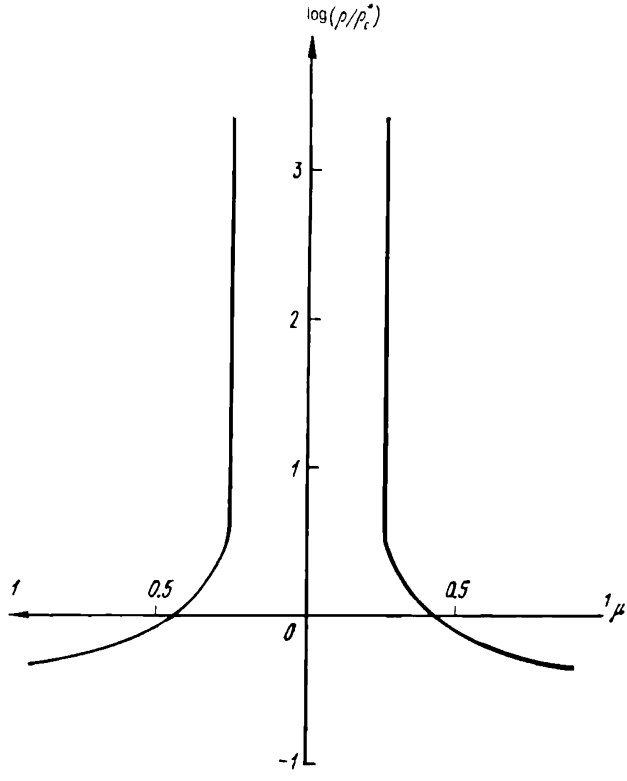


Fig. 47.—The same as Fig. 46, but after rapid cooling—at $t = 7t_c/6$, when the shock wave has just reached the Lagrangian coordinate $\mu = 0.25$. In Eulerian coordinates the region $\rho = \infty$ is compressed into a single point.

increases in the newly compressing layers proportional to μ^2, but it decreases in the matter compressed earlier by virtue of adiabatic expansion. Finally, the temperature is computed under the assumption of a rapid energy exchange between ions and electrons.

We shall now successively renounce the simplifying assumptions and shall construct a more realistic picture of the temperature distribution and of the thermal processes in the compressed gas. In the spirit of the method of successive approximations, we take the pressure to be that obtained earlier in the adiabatic picture.

First of all notice that in the original, unperturbed gas the temperature and the pressure differ from zero. Upon smooth compression to a specified pressure this gas is therefore compressed to a finite density. Correspondingly, the gas acquires a finite temperature.

Examining the thermal processes in a homogeneous Universe, we found that the thermal exchange between matter and radiation effectively ceases when, roughly, $z = 200$, $T = 540$ K, and $\rho = 0.8 \times 10^{-22}$ g cm^{-3} (for $\Omega_0 = 1$). In the subsequent evolution, regardless of whether monotonic expansion or expansion which changes into smooth contraction takes place, the pressure, temperature, and density are connected by the adiabats

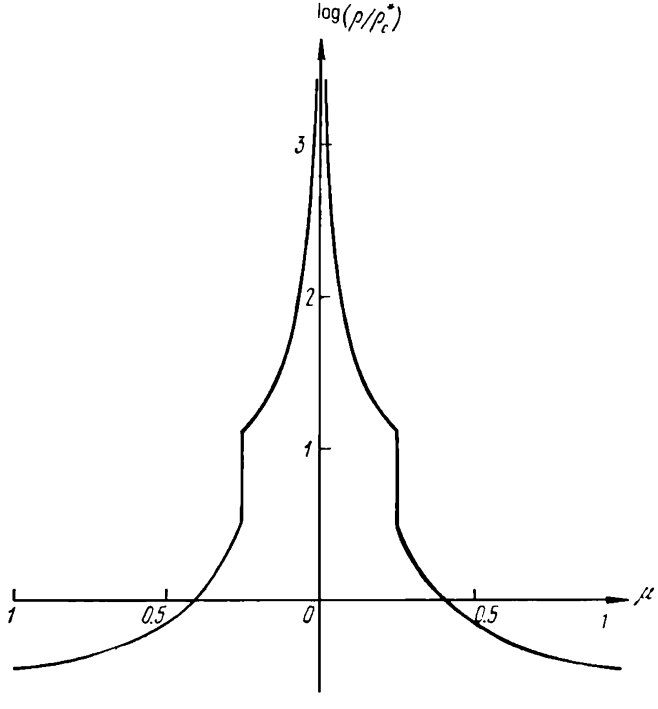


FIG. 48.—The same as Fig. 46, but for adiabatic compression. In this case, with cooling absent, the density does not become infinite after the compression.

(isentropic curves) of a monatomic gas. Thus, $\rho \sim P^{3/5}$ and $T \sim P^{2/5}$. In particular, when $z = 5$ (more precisely, when $z + 1 = 5$), we obtain $T = 540(5/200)^2$ K $= 0.34$ K for the unperturbed gas (at the same time, the radiation temperature is 14 K). Upon compression to $P = 1.44 \times 10^{-12}$ dyn cm^{-2}, which roughly corresponds to the pressure in a pancake with a mass $M = 10^{13} M_\odot$ when $z_c = 5$, we obtain $T = 400$ K and $\rho = 0.5 \times 10^{-22}$ g cm^{-3}. Compare this quantity with the average density at the same time, $\langle\rho\rangle \approx 10^{-27}$ g cm^{-3}. Roughly speaking then, one can consider the density to be $\rho_m = 5 \times 10^{-23}$ g cm^{-3} in those places where, according to the calculation, the zero-temperature density would exceed ρ_m. We thus obtain the thickness of the layer in which $\rho = \rho_m$, i.e., the thickness of an adiabatically compressed layer.

Behind the front, $\rho/\langle\rho\rangle = \mu^{-2}$. Equating $\mu^{-2} = 5 \times 10^{-23}/(1.2 \times 10^{-27}) = 3.3 \times 10^4$, we then obtain $\mu \approx 0.005$. The same result is obtained if one finds the temperature distribution in the $T_0 = 0$ approximation and then says that a temperature lower than the adiabatic temperature with the given pressure is impossible and, based upon this, replaces a part of the $T = T(\mu)$ curve by $T = \text{const.} = T_{ad}$.

The next stage consists of an account of the heat losses. The thermal radiation per unit volume (for $kT \gtrsim I_H$ with I_H the ionization energy of hydrogen, so $I_H/k = 157{,}000$ K) is given by the expression

$$W = 1.4 \times 10^{-27} n_e^2 T^{1/2} \text{ ergs cm}^{-3} \text{ s}^{-1} . \qquad (15.4.2)$$

In a gas confined by shocks the radiation of energy occurs at a constant pressure. It is therefore necessary to form the equality

$$W = -C_p \rho \frac{dT}{dt} = -5kn_e \frac{dT}{dt} , \qquad (15.4.3)$$

and to express the rate of heat transfer in terms of the pressure and temperature as

$$\frac{dT}{dt} = -a^+ P T^{-1/2} ; \quad a^+ = 1.5 \times 10^4 \text{ erg}^{-1} \text{ cm}^3 \text{ s}^{-1} \text{ K}^{3/2} . \qquad (15.4.4)$$

From this, the time for cooling to zero temperature is

$$t_T = (a^+ P)^{-1} \int_0^{T_0} T^{1/2} dT = \frac{2}{3a^+ P} T_0^{3/2} . \qquad (15.4.5)$$

Actually, the cooling proceeds only down to the recombination temperature. By this time, the gas is not in equilibrium with the thermal radiation. The recombination condition is therefore determined not by the Saha formula but rather by the Elvert formula. The temperature is then higher than when recombination proceeds under equilibrium conditions. In the approximation considered, a pressure $P = 2.6 \times 10^{-13}$ bar corresponds to compression of about half of the matter of a pancake.

For a pressure of the order of 2.6×10^{-13} bar, the corresponding number density of hydrogen atoms is about 0.2 cm^{-3}.

It is important that the recombination time depends strongly on the initial temperature. At each moment a sharp boundary is therefore found in the gas. The part of the gas with temperature $T_0 < (\frac{3}{2}a^+Pl)^{2/3} = T_1$ completely cools to the temperature 10^4 K. The other part of the gas cools little. For example, if $T_0 = \alpha T_1$, then the final temperature is $T = (\alpha^{3/2} - 1)^{2/3}T_1 = \beta T_1$, which gives $\beta = 1.5$ for $\alpha = 2$ and $\beta = 0.45$ for $\alpha = 1.2$; but $\beta = 0$ for $\alpha \leq 1$. For the parameters chosen above ($M \sim 10^{13} M_\odot$ and $1 + z_c = 5$) and at the time when half of the matter is compressed, we find $T_1 \sim 4 \times 10^5$ K.

A quite distinctive temperature distribution (Fig. 49) and corresponding density distribution (Fig. 50) are thus obtained. For the numerical calculations see Doroshkevich and Shandarin (1973, 1974). The distinguishing features of the distribution are these:

1. About 1% of the matter is subject only to adiabatic compression and has a quite low temperature.

2. A small part of the matter, 2–3%, is heated by the shock to a temperature between 10^2 K and 10^4 K. This matter is not ionized and cools slowly, only by virtue of adiabatic expansion.

3. About 20% of the matter is compressed by the shock wave and is heated to a temperature greater than 10^4 K. Thereafter, however, it cools to 10^4 K by radiating and, to a significant degree, recombines.

4. About 25% of the matter is heated by the shock to a temperature greater than 5×10^5 K (and up to $[1–2] \times 10^6$ K) and remains hot. It does not cool noticeably.

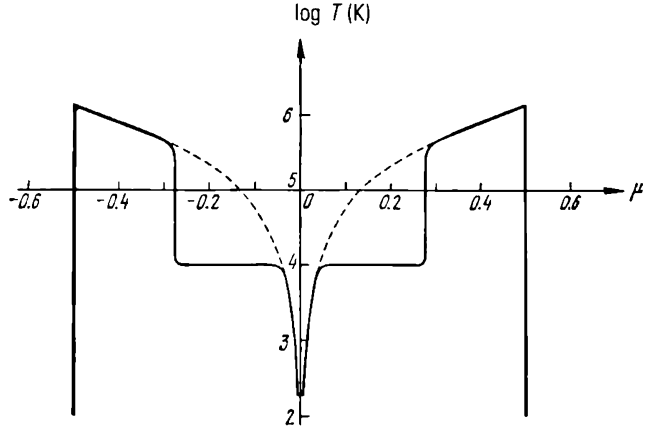


FIG. 49.—The temperature distribution in a developing cluster. The temperature T is shown as a function of the Lagrangian coordinate μ inside of the compressed matter at the moment when the shock wave is located at $\mu = 0.5$. The solid curve takes account of radiative cooling, while the dashed curve is for the case of no cooling.

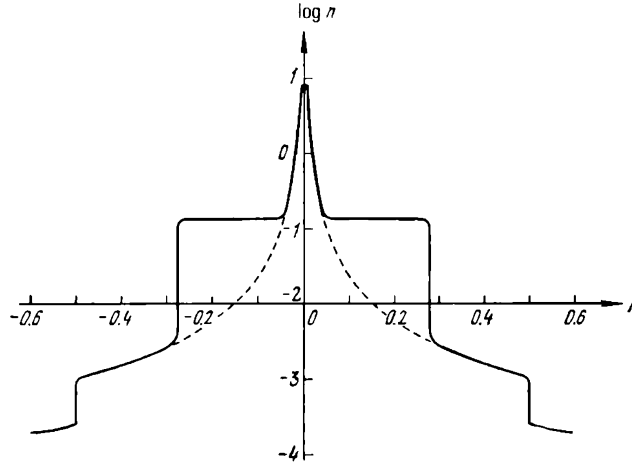


FIG. 50.—The density distribution in a developing cluster at the same moment as in Fig. 49. The number density of particles n is shown as a function of the Lagrangian coordinate μ. The solid curve takes account of radiative cooling, while the dashed curve is for the case of no cooling.

5. Half of the matter is not subject to the action of the shock at all. Indeed, it is known from observations that neutral hydrogen is practically absent from the space between clusters of galaxies. Convincing proof of this is given by the spectra of quasars with redshifts $z > 2$: neutral hydrogen would totally absorb radiation shortward of the Lyman-α line. By virtue of the redshift, this radiation falls in the region of the spectrum accessible to ground-based telescopes (Gunn and Peterson 1965). Independent of the detailed pancake theory, one cannot conceive of there being absolutely no matter in the space between galaxies. Consequently, the matter in this space is completely ionized. The suitable explanation for this ionization of the gas is one of the main problems facing cosmological theory.

It has been suggested that the radiation from the compressed gas (from the "pancakes") can ionize gas not subject to shock compression (Zel'dovich and Sunyaev 1972). This radiation is probably insufficient to do this, however. It is possible that the ionization is accomplished by the action of quasars of an earlier epoch with redshift $z \gtrsim 4$. Such quasars are not directly observable. Perhaps young, bright galaxies are important. In both cases the radiation comes from the first pancakes.

All of the calculations are made quite difficult by the indefiniteness of the most important parameters—the Hubble constant and the total matter density. In the interval from $H_0 = 100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and $\Omega_0 = 1$ to $H_0 = 50\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ and $\Omega_0 = 0.03$, the density changes by a factor of 120! The processes of ionization and maintenance of the gas in an ionized state require incomparably less radiation in the case of small H_0 and Ω_0.

Can one observe the hot compressed gas clouds—"pancakes"—directly?

That is, can one observe protoclusters in the distant past, before their transformation into present day configurations? Hardly. Estimates show that the optical and ultraviolet radiation emitted by them can at no time be observed against the background of other, nearby sources.

One can apparently observe X-rays only from the statistically rare, largest clusters. It is at the same time important that the high-temperature gas does not cool by means of radiation. If, moreover, the gas is gravitationally bound, then it does not undergo expansion and hence does not cool adiabatically. The X-rays from such sources, as the Coma cluster, are presumably the bremsstrahlung from hot gas.

The possibility of an independent (not related to the X-ray emission) determination of the amount and temperature of this gas was discussed in connection with RR. Here we are discussing the question because one can assume that the hot gas has a cosmologically significant origin. The gas is first compressed by a shock in the process of "pancake" formation. The gas then turns out to be bound when the "pancakes" transform into clusters. This is the time when galaxies and stars arise in the cold, central regions of "pancakes" and the "pancakes" become approximately spherical; the gravity acting over the surfaces collects all of the matter into clumps or clusters.

Notice now a general regularity. If the gas heats up to a temperature T_1 at a given redshift z_1 and then adiabatically expands, the hardest radiation perceptible to observation (denoted by the index 0) comes from the initial stage with $T_{\text{eff},0,1} = T_1(1 + z_1)^{-1}$. (The first subscript on T_{eff} denotes the reception redshift $z = 0$ and the second the emission redshift $z = z_1$.) A hot nonrelativistic gas expanding with the Universe cools, thus $T \sim T_1[(1 + z)\,(1 + z_1)^{-1}]^2$ later, and the gas temperature falls rapidly. The decrease of the redshift does not compensate for this fall. For radiation emitted when $z = z_i$ and received today, we find

$$T_{\text{eff},0,i} = T_{\text{eff},0,1}\left(\frac{1 + z_i}{1 + z_1}\right) < T_{\text{eff},0,1}\,. \tag{15.4.6}$$

If the gas acquires the temperature T_1 at z_1 and then neither cools nor expands (because it is gravitationally bound), then the hardest (highest energy) radiation observable arrives from a nearby object, characterized by the temperature

$$T_{\text{eff},0,0} = T_1 > T_{\text{eff},0,1}\,. \tag{15.4.7}$$

In short, research by all methods on powerful, extragalactic X-ray sources is interesting not only for itself but also for cosmology and for the theory of galaxy formation.

Finally, it is apparently possible, though difficult, to observe protoclusters because of the 21 cm radiation from the neutral hydrogen in the "pancake" region where the temperature is insufficient for ionization to

occur (Sunyaev and Zel'dovich 1972*b*, Zel'dovich and Sunyaev 1974, Novokreshchenova and Rudnitsky 1973). If one could successfully observe this radiation and demonstrate its association with a "pancake," then the redshift of the line would permit one to determine the epoch of the "pancake."

According to an estimate, the brightness temperature at the center of the line can be of the order of 1000 K. The expected spectral width of the line $\Delta\nu/\nu$ is of order 10^{-4} to 10^{-3}. The angular size of the radiation region is estimated as of the order of several minutes of arc.

The difficulty in the observations is related to the lack of knowledge of the absolute wavelength, unknown because the value of z at which the compressed gas is formed is unknown. Moreover, the direction of searches is not known beforehand. The compressed gas must be observed at a value of z significantly exceeding the redshift of "ordinary" radio sources such as quasars that radiate at all wavelengths. Soon after the formation of a cloud of compressed gas, 21 cm radio emission must cease or at least substantially weaken. This is because a significant part of the neutral hydrogen is transformed into stars and part of the remaining free gas is again ionized by the radiation from the young stars. In practically any theory, one can expect such an effect, but at quite different moments of time, or values of the redshift z. Thus, observations of the radio emission from protogalaxies would have great theoretical significance.

Also, large dimensionless numbers can occur in the relation between the masses of gravitationally bound objects (in particular, clusters of galaxies) and the quantity with the dimensions of mass—the density multiplied by the cube of one-half of the wavelength—that appears in the theory.

15.5 Cluster Masses and Protocluster Fragmentation

The estimate previously given of the mass of a protocluster that conforms to the characteristic (critical) perturbation wavelength is only a very rough approximation. In a statistical problem, moreover, it is evident that one ought to expect both a distribution of masses and a dependence of this distribution on the form of the perturbation spectrum, i.e., on the law describing the change of perturbation amplitude with wavelength. We shall in this section attempt to characterize these masses more accurately.

The second question considered here concerns the subsequent fate of the isolated bodies of gas (protoclusters) whose structure was described above. The gravitational instability of protoclusters leads to further fragmentation into different objects (galaxies and quasars).

Doroshkevich (1970) has studied the first question in detail by the methods of probability theory. He makes the natural assumption that the initial spectrum of density perturbations is of a power-law form, $|A|^2 \sim k^n$. The phases of the various waves are uncorrelated, and the amplitudes are

distributed according to a Gaussian distribution. Thus, with dissipative damping accounted for, the result

$$\Delta(M) = B(t) \exp\left[-(M_c/M)^{1/3}\right] \left(\frac{M}{M_c}\right)^{-n} e^{3n}(3n)^{-3n} \qquad (15.5.1)$$

is obtained. In accordance with the definition of Δ, this leads to

$$[\langle(\delta\rho/\rho)^2\rangle]^{1/2} = \left[\int_0^\infty \Delta^2 d \ln M\right]^{1/2} = B(t)e^{3n}\left[\frac{(6n)!}{2n(6n)^{6n}}\right]^{1/2}. \qquad (15.5.2)$$

One can make use of this equation to determine the time when nonlinear effects become substantial in a significant fraction of the matter or, more simply, when a significant part is compressed into "pancakes." For this purpose, it is sufficient in the first approximation to find the fraction of the matter satisfying $\alpha b/a > 1$ (see eq. [15.3.1]). As the hydrodynamical calculations showed for the one-dimensional problem, the amount of matter compressed by the shock exceeds the amount of matter that, by itself obeying the condition $\alpha b/a \geq 1$ (in the absence of the shock wave), approaches infinite density.

At the initial stage and given instantaneous cooling, this excess amounts to a factor of $\sqrt{3}$. Without cooling, the factor is $\sqrt{1.2} \leq 1.1$. With the help of the distribution function for the α, β, and γ probabilities, one can find $W(\alpha)$, the total probability that α is greater than a fixed value, by integrating over β and γ. This probability then applies for any value of β and γ. From this follows a law for the change with time of the total amount of compressed matter (shown in Fig. 51). We plot t/t_c along the abscissa and $2W(\alpha b/a = 1)$ along the ordinate. We are defining t_c in this statistical problem as the time when $(\delta\rho/\rho)^2 = 1$, where $(\delta\rho/\rho)$ is computed according to the linear theory. For comparison purposes, the curve relating

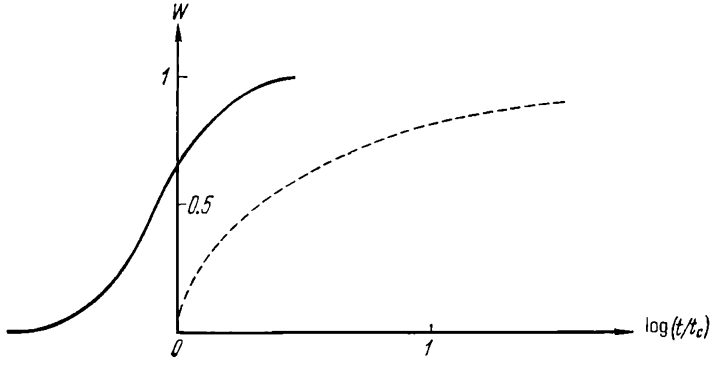


FIG. 51.—The time evolution of the fraction of the matter compressed by the shock wave. The solid curve is the result of the statistical theory, while the dashed curve is for a one-dimensional problem with a sinusoidal perturbation.

to the one-dimensional problem with sinusoidal perturbations is also presented. The dashed curve for the case of cold clouds is distinguished from the solid curves not primarily by the one-dimensional nature, but rather by the statistical nature. The calculations were performed for $\Omega_0 = 1$.

Only the absolute value of t_c depends on the initial perturbation amplitude. If the ratio t/t_c or its logarithm is plotted along the abscissa, then the curves are universal and do not depend on the amplitude.

Doroshkevich and Shandarin (1975), however, go substantially further in detailing the nonlinear stage. To the accuracy of the approximate theory they manage with the help of the correlation function to determine at each moment the density of "crystallization nuclei"—the number of high-density regions, or the number of "pancakes," per unit volume. Also determined are the volume and the mass located around a pancake which, with a high probability, are gravitationally associated with this "pancake."

The calculations are complicated, and in the late stages they become unreliable. This is because not all of the local maxima of α give independent pancakes. The particles are perhaps compressed by a shock wave of another pancake before the α-dependent infinite-density layers arise in it. It is evident that then the formulae of the approximate theory lose their validity because one can by no means neglect pressure in the compressed matter. The main reliable conclusion of the calculations is that the thickness of the typical pancake (in Lagrangian coordinates) is of order R_c while the area is $\sim 30\, R_c^2$.

The appearance of large dimensionless factors in the relevant formulae is very instructive. These factors demonstrate the inadmissibility of the simplified, naive approach to a complicated statistical problem. The numerical calculations of the three-dimensional case can rectify the first estimate; nevertheless, the incorrectness of a naive identification of M_c with the mass of clusters is quite unexpected!

The question of the fragmentation (the breakup into parts, or fragments) of a protocluster has been studied by Sunyaev and Zel'dovich (1972*b*).

The peculiarity of the problem is that one examines the gravitational instability of matter occupying a thin layer with a thickness many times less than its longitudinal size. Moreover, the dense, cold matter is subject to the external pressure of the surrounding hot matter, the density of which is much less. During the initial stage of growth of the pancake, this external pressure is greater than the gravitational forces between particles in the dense layer.

The problem of the gravitational instability of a thin (in the limit, infinitely thin) layer has been studied by many authors in connection with the theory of galactic disks. Our problem is simpler since one can assume that there is no rotation and no peculiar velocities in the plane of the disk.

Let us therefore examine a homogeneous disk of given surface density σ

(with dimensions g cm^{-2}). We seek a solution in the form

$$\frac{\delta\sigma}{\sigma} \quad (\text{const.}) \exp(\omega t + i\boldsymbol{k}\cdot\boldsymbol{x}) , \tag{15.5.3}$$

where $\boldsymbol{k}$ and $\boldsymbol{x}$ are two-dimensional. The gravitational potential is meanwhile defined in three dimensions. With due regard for the equation $\nabla^2\phi = 0$ outside the disk, we write

$$\delta\phi = (\text{const.}) \exp(\omega t + i\boldsymbol{k}\cdot\boldsymbol{x} - |k||z|) . \tag{15.5.4}$$

The amplitude of the potential is determined by the boundary condition

$$\left.\frac{\partial\delta\phi}{\partial z}\right|_{z=0^+} - \left.\frac{\partial\delta\phi}{\partial z}\right|_{z=0^-} = 4\pi G\delta\sigma \tag{15.5.5}$$

at the plane. The equations then give for the quantity ω in the law for the increase of the perturbations

$$\omega = (2\pi G\sigma|k|)^{1/2} . \tag{15.5.6}$$

The characteristic difference from the three-dimensional problem consists of the dependence of the rate of increase on the wave vector $|k|$. Recall that in the three-dimensional case,

$$\omega = (4\pi G\rho - a^2k^2)^{1/2} \longrightarrow (4\pi G\rho)^{1/2} \tag{15.5.7}$$

as $k \longrightarrow 0$ (see chap. 10). The dimensions of $\sigma|k|$ are g cm^{-3}, just as is true for ρ. It is therefore not surprising that $\sigma|k|$ enters into the formulae in place of ρ.

For a disk of finite thickness, the formulae for the gravitational potential change. The main change is that there appears a force acting against gravitational instability. This force, acting in the plane of the disk, is the product of the internal pressure and the cross-sectional area perpendicular to the plane of the disk. More precisely, it is the gradient of this quantity, and it arises because of the perturbation. Now the cross-sectional area per unit length of a line in the plane of the disk is the thickness h of the disk. Accounting then for the finite, nonzero thickness and for the pressure, we obtain

$$\omega = (2\pi G\sigma|k| - \sigma^{-1}Phk^2)^{1/2} . \tag{15.5.8}$$

Thus, for a disk of nonzero thickness, there is a critical wavelength analogous to the Jeans wavelength in the three-dimensional problem:

$$\omega = 0 \quad \text{at} \quad k = k_J = 2\pi G\sigma^2P^{-1}h^{-1} , \quad \lambda_J = 2\pi/k_J . \tag{15.5.9}$$

There is also an important quantity not analogous to anything in the three-dimensional problem. This is the wave vector k_m and the wavelength λ_m of the most "dangerous," i.e., most rapidly growing, perturbation, for which ω is a maximum. It is easy to convince oneself that $k_m = \frac{1}{2}k_J$ and $\lambda_m = 2\lambda_J$. If the pressure in the layer is determined only by the weight (through

gravity) of the matter in the layer, then $P_c = \pi G\sigma^2/2$ in the middle of the layer; at the edges, it is evident that $P_c = 0$. The average pressure is therefore of the order of $P = G\sigma^2$. Substituting this value in the expressions for λ_J and λ_m, we find

$$\lambda_J = h\,, \quad \lambda_m = 2h\,. \tag{15.5.10}$$

Furthermore, using $\sigma = h\langle\rho\rangle$, we have

$$\omega_m = 0.885(4\pi G\langle\rho\rangle)^{1/2}\,. \tag{15.5.11}$$

These results are wholly natural from the point of view of dimensional analysis.

If a layer is surrounded by a hot, light gas, then it is necessary to pay attention to this external gas pressure. One can, however, neglect its gravity and the hot gas perturbation caused by the perturbation of the potential. In this case the critical wavelength and the "dangerous" wavelength grow in the ratio

$$\frac{P_{\text{gas}} + P_{\text{grav}}}{P_{\text{grav}}}\,,$$

where P_{gas} is pressure of the gas exerted on the surface of the layer, and P_{grav} is given above as $G\sigma^2$.

In the study by Zel'dovich and Sunyaev, these considerations are applied to the disintegration of a planar protocluster—a pancake—the structure of which is outlined in the preceding section. According to the estimates, the inner, most dense (adiabatically compressed) layer must disintegrate into masses of the order of 10^7–$10^8\,M_\odot$. A layer that cools to 10^4 K disintegrates into masses of the order of 10^{11}–$10^{12}\,M_\odot$. A hot gas with temperature $T > 4 \times 10^5$ K remains gravitationally unbound in part and in part forms a hot halo around masses of 10^{11}–$10^{12}\,M_\odot$. A hypothesis has therefore been propounded according to which the most dense but small masses represent quasars and galactic nuclei.

Masses of order 10^{11}–$10^{12}\,M_\odot$ naturally are identified with galaxies. From this point of view, several general properties, connected with various types of objects are explained in a natural way.

In the adiabatically compressed gas the velocity must not have vortex components. The initial density of this gas is of the order of 30 atoms cm^{-3}. In the course of further compression, the average density can grow. After the formation of the first "pancakes," the part of the gas enveloped by the shock wave is heated to a high temperature. The hard radiation from these layers then heats the gas to a temperature of at least several thousand degrees. Such heating does not affect the general picture of the subsequent evolution of the pancakes. In the pancakes formed later, however, the middle cold layer is then not so dense and cold as was the case for the first

pancakes. On this basis, perhaps one might assume that the formation of quasars (and nuclei of galaxies?) ceases earlier than does the formation of galaxies. It is possible that the sharp fall in the number of quasars and radio sources for $z > 3$ or 4 noticed by Longair (1966; see also Doroshkevich, Longair, and Zel'dovich 1970) is related to this circumstance.

In the first pancakes about 1% of all of the matter passes, according to assumption, into quasars and galactic nuclei. Such an amount agrees with the rough estimates derived from the assumption that the active life of a quasar extends over 10^5–10^6 years. Quasi-stellar radio-quiet objects are here included among the quasars.

In every pancake with a dense central layer, one can expect the formation of about 100 dense objects. In such a case, must the quasars be arranged in groups? Must their positions not be correlated? Observations do not confirm such a correlation; indeed, such a correlation should not exist. This follows if one keeps in mind that the active period of quasars is 10^5–10^6 yr, which is many times less than both the present-day cosmological age (10^{10} yr) and the time of formation of the pancakes (3×10^8–10^9 yr). At each time we observe of the order of 10^{-3} of the total number of dense objects—potential quasars—so that the average number of observed quasars for each protocluster is significantly less than unity. Expected pairs of quasars must constitute less than 10% of the total number of them.

Turning to galaxies, notice first of all the stability of the temperature in the layer out of which the galaxies presumably form. This temperature ($\sim 10^4$ K) is determined by the features of the recombination law and of the radiation from the optically thin gas. The gas radiates most strongly when it is partially ionized. Under this circumstance, the bulk of the emission is in lines and in recombination radiation. After recombination, the emission decreases sharply. Therefore, the temperature which the gas achieves depends weakly on the history and parameters of the pancake as a whole (its mass and formation time). According to a remark made by Ozernoy, this temperature stability might well reflect the small scatter in the density of galaxies from the average density.

We here conclude the discussion of the stages of protogalaxies at the stage of fragmentation of the pancake into separate gas clouds. The subsequent development, i.e., the emergence of stars in this cloud, lies beyond the limits of cosmology.

15.6 Galaxy Rotation

Is the emergence of rotation possible in a theory in which it is assumed that the initial perturbations are vortex-free? For a long time it seemed that rotation does not arise in such a situation. Indeed, such an opinion was not entirely without foundation.

The well-known Kelvin-Helmholtz theorems say that gravitational

forces, and radiative pressure, are capable of creating during the RD stage only a potential-type velocity field, i.e., a field with vanishing vorticity. Before recombination, the density and pressure are uniquely related to each other and the forces connected with the pressure do not give rise to vortex motion. Finally, viscosity too does not give rise to vorticity in potential-type motion.

Concerning the concept of vorticity in relativistic hydrodynamics, see the work of Zel'manov (1959*b*), Vainshtein and Ruzmaykin (1972), Chernin and Eidel'man (1969, 1971), Chernin (1969), and Tiunov and Chernin (1971). Taking account of the dependence of the viscosity on density leads to the appearance of vorticity, but the effect is negligible.

In principle, the emergence of vortex motion is not excluded in the case where the potential-type motion leads to shock waves. In the theory of adiabatic perturbations, however, the amplitude of the potential-type motion during the RD era is quite small: $u/c \sim \delta\rho/\rho \sim 10^{-3}$ to 10^{-4} in the region of the spectrum corresponding to masses $\gtrsim 10^{13}\, M_\odot$. Such an amplitude does not lead to the emergence of shock waves even after a cosmologically significant time. The amplitude can be greater in the short-wave region where $M \ll 10^{13}\, M_\odot$, however. It is not excluded that these waves transform into shock waves earlier because of viscosity. This process can have significance for the appearance of entropy perturbations and will be discussed below. However, such shocks would lead to perturbations in the Rayleigh-Jeans part of the RR spectrum which are unobserved. This makes such a hypothesis unlikely. Moreover, the vorticity that is connected with small-scale shock waves must be of correspondingly small scale. Such a vorticity completely decays by recombination. If primordial vortices are absent, then the realistic action of this process in the early period of evolution is such that the motion remains vortex-free and of the potential type. Therefore we conclude that adiabatic perturbation theory probably leads to vortex-free motion at recombination.

The second side of the question concerns the relation between the body's angular momentum and its vorticity. For rotation of a solid body, the angular velocity of each volume element identically equals the angular velocity $\boldsymbol{\Omega}$ of the whole body. Therefore, $\nabla \times \boldsymbol{u} = 2\boldsymbol{\Omega}$. When $\nabla \times \boldsymbol{u} = 0$, $\boldsymbol{\Omega} = 0$ too, so the angular momentum $\mathfrak{M}$ of the body as a whole also equals zero. It is curious that in the general case of anomalous (not spherical) shapes the bodies are characterized by a moment of inertia tensor I_{ik}, while the direction of the angular momentum $\mathfrak{M}$ does not coincide with the direction of $\boldsymbol{\Omega}$ or of $\boldsymbol{\Omega} \times \boldsymbol{u}$. Since $\mathfrak{M}_i = I_{ik}\Omega_k$, one cannot represent the angular momentum $\mathfrak{M}$ as an integral over volume of $\nabla \times \boldsymbol{u}$.

For a definite mass of fluid or gas, the angular momentum of the mass as a whole cannot be written as an integral of the local vorticity. The well-known vortex-free motions of an incompressible fluid with finite angular momentum, occurring when a vessel filled with fluid rotates, dis-

proves the common belief that angular momentum is tied to vorticity. It is important, however, to note that such motion is possible only in the case of a non-axially symmetric vessel. One should not consider motion described by the velocity field $u_x = -ay/(x^2 + y^2)$, $u_y = -ax/(x^2 + y^2)$ and, $u_z = 0$ as vortex-free. The curl of the velocity, $\nabla \times \boldsymbol{u} = -4a\delta(x^2 + y^2)\boldsymbol{n}_z$ in this case. It equals zero everywhere except on the z-axis, where it is infinite.

Indeed, if the fluid has a constant density everywhere and fills an axisymmetric container, and has vanishing vorticity, then the projection of the angular momentum on the z-axis is identically zero. It is easy to convince oneself of this: find the projection of the angular momentum in the direction of $\boldsymbol{\omega}$ ("the axis"),

$$\mathfrak{M}_\omega = \rho \int (\boldsymbol{r} \times \boldsymbol{u})_\omega d^3r , \tag{15.6.1}$$

by dividing the volume under consideration into a ring with center at the axis. For such a ring,

$$d\mathfrak{M}_\omega = \rho r_1 dS \oint u_l dl , \tag{15.6.2}$$

where r_1 is the radius of the ring, dS is its cross section, dl marks off distance along the ring, and u_l is the projection of the velocity on dl. The integral along the ring represents the circulation of the velocity and consequently, according to the Stokes theorem, is expressed in terms of the integral of the vorticity over a closed surface whose boundary is the ring. For vortex-free motion,

$$\oint u_l dl \equiv 0 .$$

It is apparent from this that the problem of the emergence of angular momentum in vortex-free motion is connected with violation of axial symmetry or to the variable density. One can consider a volume of fluid confined by the vessel as the limiting case of a sharp dependence of the density on the coordinates. Inside the volume V, $\rho = \rho_0$, while outside $\rho = 0$; on the surface S bounding the volume, there is a discontinuity in ρ. We shall assume that the fluid moves in accordance with a potential but has nonzero total angular momentum on account of the asymmetric form of the volume occupied by the fluid. Let this fluid be isolated from the action of external forces, so that the angular momentum is conserved. Does the motion become vortical in the course of time?

A general property of rigid-body rotation is that it represents motion with a minimum kinetic energy for a given total angular momentum. Therefore, a body isolated from external forces tends to a state of rigid-body rotation as time passes. (Internal sources of energy such as nuclear reactions can give rise to deviations from rigid-body rotation.) But rigid-

body rotation is vortex motion. Viscosity evidently must give rise to a transition from potential-type motion with angular momentum to rigid body rotational motion; i.e., the viscosity must give rise to vorticity. It does so by transforming into heat the surplus of kinetic energy exceeding the energy of rigid-body rotation with the same angular momentum. In rigid-body rotation, moreover, the fluid does not deform.

A detailed examination shows that for a bounded body with angular momentum undergoing incompressible motion of the potential type, the entropy grows throughout the volume when account is taken of the viscosity. Vorticity arises, however, first at the boundary; gradually, after the time $(\nu L)^{1/2}$ (ν is the kinematic viscosity and L is the linear size of the body), it encompasses the whole volume. As a matter of fact, specifying $u_i = \mathfrak{D}_{ik}x_k$, where $\mathfrak{D}_{ik}$ is symmetric and trace-free, we obtain a nonzero value for the entropy increase

$$dS/dt = \text{const.} \sum_{i,k} (du_i/dx_k + du_k/dx_i)^2 .$$

On the other hand, $(du_k/dt) = \nu\nabla^2 u_k \equiv 0$. The other contributions to the acceleration du_k/dt caused by the pressure and by gravity do not produce any vorticity. This fact naturally agrees with the properties that there can be no angular momentum in an unbounded space and that viscosity does not directly generate vorticity locally.

A detailed (perhaps too detailed) discussion of the theorem of the conservation of vorticity and of the relationship between the vorticity and the angular momentum is now necessary for an understanding of the internal logic of the work of recent years. This work has elucidated the possible ways in which rotation might arise in such cosmological objects as galaxies and clusters of galaxies. We shall therefore consider three situations.

A. The Early Postrecombination Period of Small Adiabatic Perturbations

When separate perturbation modes grow in accordance with the linear theory, the vorticity, the curl of the velocity, equals zero. Calculating then the angular momentum of an arbitrarily chosen sphere in the linear approximation, with $u \neq 0$ and $\rho = \rho_0$, we obtain zero. However, if we take account of the deviation of the density from the average value ($\delta\rho/\rho \sim |u|t/l \neq 0$), then in this approximation, quadratic with respect to the amplitude of perturbations, the angular momentum of the sphere differs from zero (Peebles 1969*a;* Doroshkevich 1970). In the following development, we shall omit all dimensionless factors in order to obtain a rough estimate of the angular momentum.

For a density-perturbation amplitude of order $\delta\rho/\rho \sim 1$, the speed is $u \sim l/t$, where l is the scale over which the motion takes place and t is the cosmological time. Thus

$$\mathfrak{M} \sim \delta\rho_1 u_1 l^4 \sim \rho l^5/t . \qquad (15.6.3)$$

We then substitute $l \sim (G\pi G\rho)^{-1/2}$ to obtain

$$\mathfrak{M} \sim G^{1/2}\rho^{3/2}l^5 \sim G^{1/2}M^{3/2}l^{1/2}, \tag{15.6.4}$$

where $M = \rho l^3$. But this is just the angular momentum for which the centrifugal force keeps the mass M in a disk configuration of linear size l!

The authors mentioned above do not draw such a conclusion. Indeed, this conclusion is quite trivial since there are no dimensionless parameters in the problem. The attention of these authors is instead concentrated on a computation of the dimensionless factor in the final formula for $\mathfrak{M}$(15.6.4). According to their estimates, this factor is much less than unity ($\sim$0.005).

B. The Shock-Wave Induced Compression of the Cooled Gas at $z < z_c$ after Recombination

In the nonlinear picture (§ 15.3), the evolution of perturbations leads to the formation of strong shock waves. Given such a shock wave, conservation of vorticity no longer holds. More precisely, one must talk about the circulation of the velocity along a contour formed by certain particles. For adiabatic motion $\int \boldsymbol{u} \cdot d\boldsymbol{l}$ is conserved, but the vorticity can change. If the initial values of the vorticity and of the circulation equal zero, however, then both quantities remain equal to zero. One can thus talk about the preservation of vortex-free motion under the conditions provided in the Kelvin-Helmholtz theory. Formally, this is connected with the growing entropy of the shock wave. Thus, when after the shock wave passes the gas cools from a temperature of the order of 4×10^5 K to 10^4 K, its vorticity is not conserved.

The importance for the problem of the emergence of galaxy rotation of the nonconservation of vorticity in a shock wave was noticed by Chernin (1970) and was then studied in detail by Doroshkevich (1973). Very roughly, imagine a shock wave that, in the first approximation, coincides with the plane $x = 0$. Gas falls, i.e., $u_x < 0$ at $x = 0$, but without moving strictly along parallel lines, so $u_y \neq 0$ and $\partial u_y/\partial x \neq 0$. Nevertheless, it does move in a vortex-free manner, so $\partial u_y/\partial x = \partial u_x/\partial y$ and $\partial u_z/\partial x = \partial u_x/\partial z$. The shock wave compressed the gas by a factor 4; therefore the motion parallel to the x-axis is proportionately slowed so that in the compressed gas

$$u_x^{\text{after}} = \tfrac{1}{4} u_x^{\text{before}}$$

and

$$\frac{\partial u_x}{\partial y} = \frac{1}{4}\frac{\partial u_x}{\partial y}\bigg|_{\text{before}}, \qquad \frac{\partial u_x}{\partial z} = \frac{1}{4}\frac{\partial u_x}{\partial z}\bigg|_{\text{before}}$$

behind the wave front. Assuming that the tangential velocity is conserved, we shall consider that the density increase in the shock by a factor of 4 is accompanied by a decrease in the x-component of the distance between particles. The tangential velocity remaining unaltered, we obtain $(\partial u_y/\partial_x)_{\text{after}} = 4(\partial u_y/\partial_x)_{\text{before}}$. Therefore, the compensation effect expressed by the equa-

tion $\partial u_y/\partial x - \partial u_x/\partial y = 0$ that leads to vortex-free motion before the shock-wave induced compression is completely absent after the compression. A rough approach then gives the estimate $\nabla \times \boldsymbol{u} = H\rho/\langle\rho\rangle$, where H is the Hubble parameter at the time of compression. The estimate is based on the nonlinear compression that occurs when the gradients of the velocity perturbations in one direction locally exceed the Hubble parameter, which determines the general expansion rate. As an example, if $z + 1 = (z_c + 1)/1.5 = 3$, $H = 3^{3/2} H_0 = (1.5\ 10^9\ \text{yr})^{-1}$, $\rho = 10^{-24}\ \text{g cm}^{-3}$, and $\langle\rho\rangle = 3 \times 10^{-28}\ \text{g cm}^{-3}$, then $\nabla \times \boldsymbol{u} = (7.5 \times 10^5\ \text{yr})^{-1}$.

Now let us assume that a galaxy with such an angular velocity rotates rapidly. We would obtain

$$\tau = \frac{2\pi}{\Omega} = \pi(|\nabla \times \boldsymbol{u}|)^{-1} = 2.3 \times 10^6\ \text{yr} \qquad (15.6.5)$$

for the period of rotation, which is 100 times less than what is observed!

It is possible that there are small dimensionless numbers in the exact theory. For example, the small quantity μ, the fraction of the matter included in galaxies, seemingly enters the theory. It therefore is reassuring that the order-of-magnitude calculation leaves a large margin for a decrease. Notice in addition that the central part of the pancake, which earlier undergoes adiabatic compression, remains vortex-free in this formulation.

C. Gravitational Interactions of Discrete Bodies

Now let us examine the last stage, when the perturbations lead to the formation of discrete bodies, spatially separated from one another. These bodies then interact gravitationally with one another.

The gravitational interaction leads first of all to the exchange of momentum—to a change in the direction and speed of motion. This is how two point particles and spherical bodies interact. Of course, the external gravitational fields of the latter do not differ from the fields caused by point particles. An irregular shaped body, however, has a nonzero quadrupole moment (and higher moments too) in its mass distribution. The external field of such a body has a part depending on angle:

$$\Phi_Q = \frac{\text{const.}}{3r^3} P_Q(r, \theta, \phi)\ , \qquad (15.6.6)$$

where P_Q is the quadrupole spherical harmonic. In such a potential the orbital angular momentum of a second body is not conserved. Correspondingly, in accordance with the law of equality of action and reaction (or according to the law of the conservation of total momentum), the angular momentum of the first body, the one which causes the potential Φ_Q, also changes. This change is due to the action of the field of the second body, a field we have not written out.

It is characteristic, and indeed inevitable according to potential theory,

that the quadrupole interaction decreases with increasing distance faster than does the potential from a point mass. Therefore, the theory which describes the gain of angular momentum in collisions differs sharply from the theory describing the redistribution of velocity in translational motion. For angular momentum gain, the closest encounters dominate.

Now the torque acting on a given body (with mass M_1 and quadrupole moment Q_1) due to a second body (with mass M_2) is of the order of

$$\left|\frac{d\mathfrak{M}}{dt}\right| \sim \frac{GQ_1M_2}{r_{12}^3}. \tag{15.6.7}$$

For passage with the impact parameter b^*, we have effectively that $r_{12} \sim b^*$ and $t \sim b^*/u$, where u is the speed. Thus,

$$|\Delta\mathfrak{M}| \sim \frac{GQ_1M_2}{(b^*)^2u}. \tag{15.6.8}$$

With chaotic collisions, the squares of statistically independent increments in the angular momentum accumulate. Since the frequency of collisions with impact parameter between b^* and $b^* + db^*$ equals $d^2W = 2\pi n_2 u b^* db^* dt$, we obtain

$$(\Delta\mathfrak{M})^2 = G^2Q_1^2M_2^2\frac{n_2t}{u}\int_{b_0}^{\infty}\frac{2\pi b db}{b^4} \approx \frac{tG^2Q_1^2M_2^2n_2}{b_0^2u}. \tag{15.6.9}$$

The result depends in an essential way on the choice of the lower limit of integration b_0.

A very rough estimate applicable to the process of separating galaxies is found by substituting $M_1 \approx M_2 = M$; $Q_1 = MR^2$; $n = R^{-3}$; $b_0 = R$; $t = R/u$; $u = (GM/R)^{1/2}$; and $\mathfrak{M}_0 = 0$. The result that follows, namely

$$\Delta\mathfrak{M} = (GM^3R)^{1/2}, \tag{15.6.10}$$

is evident from dimensional considerations. This is just the angular momentum necessary in order that the centrifugal force of mass M and radius R counterbalance the force of gravitational attraction.

A final question arises: can one assume that isolated bodies (protogalaxies) possess a quadrupole moment $Q \neq 0$ initially, before they gain angular momentum? This is an important question, for an isolated body with no angular momentum assumes a spherical shape, with $Q = 0$, in equilibrium.

At the time of isolation, however, the bodies are still unsuccessful in taking the equilibrium form. This is the first reason why $Q \neq 0$. In addition, even if the first body is spherically symmetric before the approach of a second body with mass M_2, then the very approach of this mass gives rise to a tidal deformation of the first body and therefore to the appearance of a quadrupole moment. In order of magnitude, this "induced" moment is less than $Q_1 = M_1R^2$ by the ratio $(M_2R^3)/(M_1b^3)$. This condition holds at the time of closest approach, when $r_{12} = b$. To achieve an angular-momentum

gain, it is still necessary that the quadrupole moment induced in the direction of the principal axes lag behind the line joining the centers of gravity of the bodies. This circumstance leads to an additional factor (R/b) or $(R/b)^2$ for $b \gg R$. After substitution of the characteristic values of b, of u, and of other quantities, however, one again finds that the resulting expression for $\langle \mathfrak{M} \rangle$ turns out the same.

Thus, only a computation of the dimensionless factors—a careful compilation of all of the powers of 2, π, . . .—can give a final answer to the question of the average angular momentum of galaxies and of clusters of galaxies in the theory of adiabatic perturbations. But a computation of this sort by "theoretical" (read this as analytic) methods is hardly possible. The formulation of the initial conditions is itself statistical, so the answer too must be formulated statistically and must appear in the form of distribution functions for the isolated bodies with respect to mass and angular momentum.

It is inevitable that numerical calculations of the three-dimensional motion will be carried out for a collection of situations with random initial data. But such a program will require several years if due regard is made for thermal phenomena. At the present time, meanwhile, one can state there are at least three different processes in principle that lead to angular momentum gain by isolated bodies. One cannot, however, maintain that these processes lead to a sufficiently large gain of angular momentum. But one can in no way maintain the converse either. That is, one cannot maintain that the theory of adiabatic perturbations is incapable of explaining the observed rotation of galaxies.

15.7 Galactic Magnetic Fields

The origin of a galaxy's magnetic field has been elucidated only in the last few years—in connection with the advances in the theory of the dynamo effect, the strengthening of a field caused by plasma motion. It is assumed that a quite weak magnetic field arises as a consequence of the rotation of the gas cloud, or protogalaxy, on account of the mass difference between protons and electrons and the difference between the interactions of electrons and protons with RR.

The subsequent intensification is related to the specific motion of the rotating gas. In this formulation, we rest on the observationally established rotation of spiral and elliptical galaxies. The choice of one theory or another to describe the origin of the rotation itself has little effect on conclusions concerning the magnetic field.

Several years ago it seemed that there were insurmountable difficulties with the theory of the dynamo effect. This led to a proposal according to which there exists a specified type of singularity and a comparatively strong ($\sim 10^{-9}$ gauss at the present time), primordial, cosmological mag-

netic field. This magnetic field is presumably frozen in the relic plasma. A galactic field ($\sim 10^{-6}$ gauss) is obtained in this formulation by compression of the magnetic lines of force when the dilute gas condenses into galaxies. In the space between galaxies, where the gas remains, an ordered field of strength $\lesssim 10^{-9}$ gauss must remain.

Today this formulation, although not refuted, appears artificial. Moreover, data concerning the ordered cosmological field are unreliable (Kawabata, Fujimoto, Sofue, and Fukui 1969; Reinhardt and Thiel 1970; Komberg and Ruzmaykin 1972). Therefore, we will only discuss cosmological models with a primordial field briefly—and then only in the last part of the book—in connection with the general physical question of the influence of particles and fields on the singularity.

Let us turn here to theories describing the generation of fields within the confines of a hot model of the Universe with no primordial field, but with small density and velocity perturbations superposed on a homogeneous and isotropic Friedmann solution.

Harrison (1968, 1970*b*) has examined the rotational perturbations occurring during the RD stage. Recall that the conservation of angular momentum for a sphere of radius R yields the condition

$$\mathfrak{M} = I\omega = (8\pi/15)\rho R^5\omega = \text{const.} \qquad (15.7.1)$$

In the course of the expansion, the radius R grows; meanwhile the total density falls as $\rho \sim R^{-4}$ since, in this period, the photon density dominates. The angular velocity therefore decreases as $\omega \sim R^{-1}$ while the linear velocity $u = \omega R$ remains constant. The matter density, taken separately, decreases as $\rho_m \sim R^{-3}$. By virtue of angular momentum conservation, therefore, as the expansion proceeds, the matter alone "would like" to rotate more slowly, with $\omega \sim R^{-2}$ and $u \sim R^{-1}$.

Under these conditions, the light electrons are dragged along by the radiation, but the heavy protons fall behind the rotation of the RD plasma. If the electrons and protons move at different velocities, then there arises an electric current with current density $\boldsymbol{j} = en(\boldsymbol{u}_p - \boldsymbol{u}_e)$, where $n = n_e = n_p$ is the particle number density, e is the elementary charge, and $\boldsymbol{u}_p$ and $\boldsymbol{u}_e$ are velocities. Such a current gives rise to a magnetic field $\boldsymbol{H}$ directed along the rotation axis. As we proceed to calculate the intensity of the magnetic field, it is necessary to remember that a change in the magnetic flux induces an electromotive force that tends to equalize the electron and proton velocities. In the first approximation, we assume that the proton velocity differs little from the electron velocity. We shall then find the values of $d\boldsymbol{H}/dt$ and $\boldsymbol{H}$ necessary so that the protons do not lag behind the electrons. Then, we shall verify that the difference between the velocities $\boldsymbol{u}_p - \boldsymbol{u}_e$ required for the creation of such a field is actually small.

In the following basic equations of motion for the protons the electric field E is defined in a reference system at rest relative to the distant, non-

corotating, parts of the plasma; we are interested in the tangential component of the electric field $E = E_\phi$:

$$u_p = \text{const.}\,; \qquad m_p \frac{du_p}{dt} = -m_p \frac{u_p}{R}\frac{dR}{dt} + eE\,; \tag{15.7.2}$$

$$E = \frac{m_p}{e}\frac{u_p}{R}\frac{dR}{dt}\,.$$

We then find that

$$\frac{d}{dt}(\pi H R^2) = -c(2\pi RE) = -\frac{2\pi m_p c}{e} u_p \frac{dR}{dt}\,. \tag{15.7.3}$$

We then substitute $u_p = \omega R = \text{const.}$ to obtain

$$\frac{d}{dt}\left[R^2\left(H + \frac{2m_p c}{e}\omega\right)\right] = 0\,. \tag{15.7.4}$$

The asymptotic solution to this equation is

$$H = -2m_p c\omega/e\,. \tag{15.7.5}$$

In the vortex theory, motion is examined for which the scale of the largest vortex equals the product of the velocity and the cosmological time. This is also the condition for the realization of turbulence. Thus, $\omega \sim 2\pi/t$, where t is the cosmological time, which at recombination (or at the end of the RD period) is of the order of 3×10^{12} s. With this value, we find that $H = 0.4 \times 10^{-15}$ gauss when $\rho_m \approx 3 \times 10^{-20}$ g cm^{-3}. The density of matter in galaxies is less than that by $\sim 10^4$. In other words, the expansion continues after recombination. If the expansion takes place in a similar way subsequently, then the frozen-in condition—the conservation of the magnetic-field flux—leads to a later decrease of H by an additional two orders of magnitude to a value of 2×10^{-18} gauss.

Mishustin and Ruzmaykin (1971) have studied the case when the rotation arises at a stage during which the photon density is small. They studied the case of nonequilibrium secondary ionization after recombination. In the first approximation, the protons and electrons rotate at the same velocity, which, for a gravitationally bound gas cloud, is time independent. However, the electrons slow down as a result of friction with the photon background. The condition from which the magnetic field strength is determined is that the induced electromotive force compensate for the friction on the electrons:

$$m_e \frac{du}{dt} = -\frac{4}{3}\frac{\sigma_T \epsilon_\gamma u}{c} - eE = 0 \tag{15.7.6}$$

Here $\epsilon_\gamma = 4 \times 10^{-13}(1+z)^4$ ergs cm^{-3} is the radiation energy density and σ_T is the Thomson scattering cross section. The induction equation has the same form as it had earlier. After a time of the order the cosmological

time, a magnetic field of the order of

$$H \sim \frac{t_{\text{cosm}}}{\tau_{\text{rot}}}(1+z)^{5/2}\left[\frac{4\pi\sigma_T\epsilon_\gamma}{e}\right] \tag{15.7.7}$$

is acquired. Here, t_{cosm} is no greater than the present cosmological time ($\sim 10^{11}$ yr) and τ_{rot} is the rotation period of a gas cloud, which, for our Galaxy, is of the order of 2×10^8 yr. The expression above does not take into account any change in radius of the gas cloud. By substituting the values $1 + z = 5$ and $\tau_{\text{rot}} \approx 2 \times 10^8$ yr, we obtain $H \approx 10^{-23}$ gauss. Curiously, both of the mechanisms set forth give a quantity independent of the density of current-carriers.

Therefore the field due to rotation turns out to be quite small in comparison with that observed in the Galaxy ($\sim 10^{-6}$ gauss). The same is the case for other generation mechanisms not considered here. This is true, for example, for the mechanism based on a combination of thermoelectric forces, convection, and rotation (Biermann 1950; Schlüter and Biermann 1950).

A satisfactory solution of the galactic magnetic-field problem is thus impossible without making use of the dynamo effect, which can produce exponential growth of a field with time. Simple symmetric motions of gas with a frozen-in magnetic field are capable of intensifying the field, but this intensification either is related to a decrease in the scale size of the field (for chaotic motion) or merely gives an effect which grows linearly with time. As an example, consider a disk with an initial radial field (except in the central region), a field whose lines are closed by returning above and below the disk through the vacuum. The field as a whole is quadrupole in nature. For differential rotation, that is, at an angular velocity dependent on the radius, the following equation first written down by Elsasser (1950) holds:

$$\frac{\partial H_\phi}{\partial t} = R\frac{\partial \omega}{\partial R}H_R\,. \tag{15.7.8}$$

Since, however, H_R remains constant, H_ϕ only grows linearly. Such growth is clearly insufficient to produce the fields observed today.

Exact theorems on the impossibility of a dynamo have been proved for axisymmetric and planar motion. For a long time, solutions were known in the three-dimensional case for which the growth of the field is exponential, but with a characteristic time that corresponds to the decay time for a frozen-in field. In other words, this characteristic time is proportional to the conductivity (a "slow dynamo"). For a dilute plasma and cosmological scales, this time is unacceptably large. One would thus think that there is a general law forbidding rapid growth. In the mid-1960s, however, a simple example illustrating rapid growth was constructed. Imagine a torus with cross section S and radius R filled with a field H_0. Now, while preserving its volume, stretch the torus to a radius $2R$. The cross section decreases to

$S/2$, but the field grows to $2H_0$ because of the conservation of flux. Now bend the torus into a figure eight and then fold the two rings of the figure eight into one torus. We obtain again a torus with radius R and cross section S. The magnetic field, however, becomes $2H_0$; the magnetic field is more intense despite the fact that the linear dimension is unchanged. To be precise, if the lines of force are at first closed circles, then the lines of force return to a given point only after two revolutions. These "subtle" properties of the field are not important, however. The proposed operation doubles the magnetic flux for each cycle of motion, and one can carry out such cycles an unlimited number of times. For a constant rate—one cycle for each time τ—we find exponential growth of the form

$$2^{t/\tau} = \exp\,[t(\ln 2)/\tau]\,.$$

The motion, described, is substantially nonplanar. There is necessarily a spiral turn present that transforms the torus into a figure eight. A theory of turbulent spiral motion with $\langle \boldsymbol{u}\cdot\nabla\times\boldsymbol{u}\rangle \equiv \alpha \neq 0$ (the turbulence is not reflection invariant) is constructed in the work of Steenbeck (1963) and of Steenbeck, Krause, and Rädler (1966). It is also shown there that the equation

$$\frac{\partial \boldsymbol{H}}{\partial t} = \nabla \times (\alpha \boldsymbol{H})\,, \tag{15.7.9}$$

where we have only written the main term, arises for a magnetic field in this situation. Notice that this yields an exponentially growing solution. These authors considered the case of low conductivity. Earlier Vainshtein (1971) showed that such motion also gives exponential growth in the case of high conductivity. A combined solution, including ordered motion and turbulence, is also possible. Then differential rotation gives rise to a nonzero H_ϕ from H_R as described above. Meanwhile, the turbulence, described by the coefficient $\alpha \neq 0$, gives rise to a nonzero H_R from H_ϕ so that there is a second equation, first written down by Parker (1955) in connection with the solar magnetic field:

$$\frac{\partial H_R}{\partial t} = \text{const.}(\alpha H_\phi)\,. \tag{15.7.10}$$

Such a system of two equations has an exponential solution (Fitremann and Frisch 1969; Parker 1971; Vainshtein and Ruzmaykin 1971).

An absolute upper limit on the rate of growth of the field is $\exp(\omega t)$, where ω is the angular velocity. With $\omega = 2\pi/\tau_{\text{rot}} \approx 3 \times 10^8\,\text{yr}^{-1}$ and $t \approx 10^{10}\,\text{yr}$, such a law would give an increase of e^{300}, which is more than sufficient. It is not easy to give more realistic estimates, though. What is necessary is growth "only" by a factor of 10^{10}–10^{20}.

As in many other aspects of cosmology, one must state that there is here no reliable theory at all. But neither is there a blind alley; not a clear, negative answer which would require the introduction of an intense pri-

mordial field. For a detailed review of the state of the theory of the dynamo effect, circa 1972, see the article by Vainshtein and Zel'dovich (1972).

As a "reserve possibility," the assumption has been advanced that a magnetic field is created by a dynamo mechanism in more dense regions, where the rotation rate is greater and the characteristic time is smaller. The magnetic field subsequently spreads out over large regions (e.g., galaxies) along with the ejected gas. The dense regions are perhaps stars or galactic nuclei (Hoyle 1969; Bisnovatyi-Kogan and Vainshtein 1971; Bisnovatyi-Kogan, Ruzmaykin, and Sunyaev 1973). The ejected field can also serve as a "seed" for the action of the "dynamo" effect.

15.8 The Theory of Entropy Perturbations

In the theory of the hot Universe, there exists a particular type of perturbation which, in principle, can lead to the formation of discrete, isolated, gravitationally bound bodies at the present time, even in the complete absence of metric, density, and velocity perturbations at early stages, and, in particular, near the singularity. The general character of these so-called entropy perturbations was discussed in § 11.4 and § 12.6. Recall a basic conclusion. Entropy perturbations with a wavelength corresponding to a mass $M > 10\,M_\odot$ remain practically constant right up to recombination. After recombination, during the neutral-hydrogen era, the entropy perturbations transform into density perturbations. Later, they become growing, potential-type perturbations if $M > M_J \approx 5 \times 10^4\,M_\odot$ and decaying sound waves if $M < 5 \times 10^4\,\Omega_0^{-1/2}\,M_\odot$. Here, the possible role of entropy perturbations in the formation of the present-day structure of the Universe will be examined.

One can make the extreme assumption that in the beginning, near the singularity, only entropy perturbations occur. At first glance, such a hypothesis leads to consequences not differing from those of the hypothesis of adiabatic perturbations. This is true because one can select a spectrum of entropy perturbations or a spectrum of adiabatic perturbations (during the RD stage) in such a way that they give the same spectrum of growing, potential-type perturbations at late stages after recombination. However, the actual profound difference between the two hypotheses and the two types of perturbation (entropy and adiabatic) is here hidden behind the assumption of the selection of an initial spectrum.

Now there exists the notion of a "natural" initial spectrum. Since there is no fundamental theory of the initial perturbations, it is not surprising that this notion is somewhat nebulous. Indeed, a definition of a "natural" spectrum is currently a statement of what it is not. It must not have such peculiar features as delta functions, discontinuities, kinks, maxima, or minima at a time so close to the singularity that all wavelengths of interest to us are much larger than the horizon size.

One can consider as "natural" the appearance of peculiar features on special scales resulting from some physical processes occurring during later stages. Examples of such special scales are that corresponding to the horizon size at recombination where $M \sim 10^{16}$–$10^{17}\, M_\odot$, that corresponding to the Silk decay length, where M is of the order of $10^{13}\, M_\odot$, and that corresponding to the Jeans length for neutral hydrogen, where M is of the order of $10^5\, M_\odot$.

The coincidence of peculiar features of the primordial spectrum with characteristic lengths appearing (becoming manifest) during a later stage, would be a chance occurrence and would seem artificial. It would consequently contradict the notion of a "natural" primordial spectrum.

For adiabatic perturbations, therefore, a natural spectrum has been considered to be one not having kinks for $M > 10^{12}\, M_\odot$. After recombination, peculiarities do appear in such a spectrum. A maximum that is connected with damping appears in the vicinity of $10^{13}\, M_\odot$. Also, zeros which are substantial for $M \sim 10^{18}$–$10^{19}\, M_\odot$ arise from the Sakharov oscillation.

For the entropy perturbations considered here, the natural spectrum is one not having kinks for $M > 5 \times 10^4\, \Omega_0^{-1/2}\, M_\odot$. After recombination, the single peculiarity in the spectrum of growing perturbations turns out to be the maximum around $M \sim 10^5\, M_\odot$. This maximum is formed because the rate of growth of the perturbations goes to zero at the neutral-hydrogen Jeans wavelength.

What is the subsequent fate of the perturbed matter?

If the initial amplitude of the perturbations is sufficiently large, then they manage to grow to a point where $\delta\rho/\rho \sim 1$ at some $z > 1$. In other words, the perturbations become large prior to the present era.

For the nonlinear stage, pressure plays a characteristically important role. Remember that it is just the gas pressure that determines the size of the Jeans mass. For the perturbations under consideration, therefore, the pressure radically affects the nonlinear stage too. The approximate nonlinear theory set forth above (in § 14.2) is not applicable here. There will be no pancakes!

Let us recall the description of the evolution of the homogeneous plasma given in chapter 8. When $z \approx 1400$, $T \approx 4000$ K and recombination takes place. But for a long time afterwards, or until the time when $z \approx 200$ and $T \approx 540$ K, thermal contact between the radiation and the matter is preserved. During this time, $T_m = T_r = 2.7(1 + z)$ K. During the time when $T_m \sim (1 + z)$, the Jeans mass remains constant.

If perturbations with $M \gtrsim M_J$ grow to values such that $\delta\rho/\rho \sim 1$ in this period, then the pressure turns on immediately and halts the perturbation growth. At this time, $\delta\rho$ is only several times larger than the average gas density, which is of the order of 10^2 atoms cm^{-3} or 10^{-22} g cm^{-3} when $z = 200$.

Such perturbations turn out to be gravitationally bound. Subsequently,

the surrounding gas expands in accordance with the general Hubble expansion, but the gravitationally bound bodies, as they cool by emitting radiation, contract. Under these conditions, when even a small contraction leads to a rise in the pressure because of the adiabatic growth law for the pressure ($P \sim \rho^{5/3}$ for a monatomic gas), thereby halting the contraction, one can expect the formation of dense spherical clouds. These clouds can be somewhat flattened if the original perturbation acquires angular momentum. However, the pressure, which tends to give a body a spherical shape, simultaneously inhibits any gain of angular momentum.

What is the subsequent fate of such clouds? What is their role in the formation of the general structure of the Universe?

Attempts were made in work by Doroshkevich, Zel'dovich, and Novikov (1967*a*) to derive the entire structure of the Universe as an outgrowth of the type of perturbation described above. The hypothesis was put forward that the dense clouds transform into supermassive stars that rapidly—perhaps explosively—consume their nuclear fuel. Now the nuclear energy derived from burning hydrogen into helium, about 7 MeV per proton, is 10^5 times greater than the energy required to ionize hydrogen and heat it to a temperature of 10^5 K. Consequently, a supermassive star can heat up and perturb a mass of gas much greater—perhaps 10^4 times greater—than the mass of the supermassive star itself. It may then be sufficient that a fraction 10^{-4} of all of the matter "in the tail of the Gaussian distribution" transforms into supermassive stars and burns up in order for the remaining matter to have its properties radically changed. Further processes can then proceed under the decisive influence of the heating and perturbing of the supermassive stars. The authors hoped for growth of the characteristic scales that would approach the characteristic size of clusters of galaxies, which have masses of the order of $10^{13}\ M_\odot$. For discussion of the explosion of supermassive stars see Bisnovaty-Kogan (1968) and Fricke (1973).

Dicke and Peebles (1968; see also Peebles 1967*b*) suggested another hypothesis. They assumed that entropy perturbations are responsible for the creation of the so-called globular clusters (*amas globulaires* in French). Among the constituents of our Galaxy, observers have separated out this particular type of star cluster, which consists of 10^5–10^6 stars similar in luminosity and mass to the Sun, so that the mass of a globular cluster is of the order of 10^5–$10^6\ M_\odot$. These clusters differ from other objects, however, in their large density. Their size is of the order of 50 pc, so the average density is $\sim 10^{-22}$ g cm^{-3}.

The uniformity of the density of different globular clusters is striking. Even globular clusters belonging to different galaxies have approximately the same density. Globular clusters also stand out because of their small content of elements heavier than hydrogen. This is ascertained through the spectra of stars comprising the globular clusters. Finally, it is noticed that

in our Galaxy globular clusters are spherically distributed. Their average distance from the plane of symmetry of the Galaxy is about 5000 pc, while that of young open clusters is 120 pc.

All of these properties suggest that globular clusters are ancient constituents of galaxies.

The Dicke and Peebles hypothesis leads to an immediate estimate characteristic of the mass of a globular cluster. As was noted above, the Jeans mass for neutral hydrogen equals $5 \times 10^{+4}\Omega_0^{-1/2} M_\odot$ during the period when the temperature of the hydrogen equals the temperature of the radiation. The mass for which the perturbations are most numerous at a given $z < 1400$ is evidently greater than the Jeans mass. The specific number depends on the spectral index and, moreover, on just what factors appear when passing from the perturbation spectrum to the spectrum of masses of discrete bodies. Elaborating on this breakthrough, the authors reached the conclusion that a significant fraction of the matter (about half) can divide into gravitationally bound bodies with masses of the order of 10^5–$10^6 M_\odot$. A part of these bodies transform into globular clusters—i.e., divide into stars that remain gravitationally bound to one another—but a larger part is dispersed by means of collisions and tidal interactions. In our Galaxy, globular clusters constitute about 0.1–1.0% of the total mass.

Recently, Ruzmaykina (1972) investigated the fate of a spherical gas cloud. The situation turned out to be quite complicated, depending on a large number of processes such as the formation of H_2 molecules, and radiation from them. In the possible subdivision of a gas cloud with a mass of 10^5–$10^6 M_\odot$ into stars, thermal instability plays an essential role. Along with the transformation of the cloud into a globular cluster or an exploding supermassive star, Ruzmaykina mentions the possibility of direct collapse and the emergence of a black hole. In this involved situation, though, one ought to consider this work as interesting but not definitive.

In the scheme of Doroshkevich, Zel'dovich, and Novikov (1967*a*), an attempt is made to link the early, primary formation of initial supermassive stars with the breakup of matter into more prominent units. In the hypothesis developed by Dicke and Peebles, however, it is tacitly assumed that these two processes are not connected with each other. Some (the adiabatic) perturbations have a large scale and give rise to galaxy clusters and galaxies. In the first approximation it indeed makes no difference whether these perturbations act on a gas of discrete atoms or on a gas whose "atoms" are clouds with a mass of about $10^5 M_\odot$ each. In reality, however, the detailed picture changes substantially.

In the absence of entropy perturbations, the gas gives a structure similar to that described in § 15.2. If the gas collapses into clouds, these clouds can pass by one another without forming a dense, cold layer. A detailed theory with the two forms of initial perturbations (adiabatic and entropy) has still not been worked out.

By means of a gently sloping spectrum, one can in principle specify entropy perturbations only. One can then select the spectral index so that on a scale of order $10^6\,M_\odot$ separation occurs when $z \sim 100$, for example, while on a large scale such as $10^{13}\,M_\odot$ or $10^{14}\,M_\odot$ the perturbations are of order unity when $z = 3$. The difficulties in such a formulation arise upon extrapolating the gently sloping spectrum into the region of still longer wavelengths. This side of the matter is discussed in the next chapter in connection with the observations of RR.

In this regard there is an important paper of Gott and Rees (1975). They assume a power law spectrum of purely entropy perturbations:

$$\Delta \equiv \frac{\Delta M}{M} = 10^{-4}\left(\frac{M}{M_J}\right)^{-1/3} \sim M^{-1/3}\,, \tag{15.8.1}$$

where $M_J \approx 10^{17}\Omega_0^{-2}\,M_\odot$ is the Jeans mass prior to decoupling. Using the growth law $\Delta \sim (1+z)^{-1} \sim t^{2/3}$, they obtain a gradual increase in the mass of gravitationally bound objects M_b,

$$M_b \sim (1+z)^{-3}\,. \tag{15.8.2}$$

The density in the bound objects is proportional to the average density at the time the object became bound. Since $\langle\rho\rangle \sim (1+z)^3$, the characteristic size R of objects formed at each z is

$$R \sim (M_b/\langle\rho\rangle)^{1/3} \sim (1+z)^{-2} \sim \langle\rho\rangle^{-2/3}\,. \tag{15.8.3}$$

The first bound objects form immediately after decoupling. The power of the spectrum (i.e., $\Delta \sim M^{-1/3}$) is chosen to obtain agreement with the covariance function data of Peebles.

The consequences of a possible large amplitude for the entropy perturbations have been discussed above. Now let us pose the question of whether a theory is possible in which entropy perturbations do not play a significant role. It is clear that under such circumstances one would have to explain globular clusters in some other way.

The other side of the matter is that adiabatic perturbations occurring during the early, RD stage necessarily create some entropy perturbations. This can be demonstrated by using the methods of § 14.3. Recall the results. There are two mechanisms: the dissipation of energy, for which $(\delta\rho/\rho)_e \approx (\omega/t)_s(\delta\rho/\rho)_{ad}$, and the action of shock waves, for which $(\delta\rho/\rho)_e \approx \text{const.}\,(\delta\rho/\rho)_{ad}{}^3$. For the wavelengths corresponding to galaxy clusters, with a mass $M \sim 10^{13}\,M_\odot$ and a corresponding amplitude of $(\delta\rho/\rho)_{ad} \approx 10^{-3}$, both mechanisms give insignificant values of $(\delta\rho/\rho)_e$. Thus, the result is determined by how the spectrum is extrapolated from the initial adiabatic perturbations into the region of small masses. A theory with negligible entropy fluctuations is therefore entirely possible; it would not contradict the general principles.

15.9 The Vortex Theory

The vortex theory derives from the assumption that during an early RD stage the plasma is in a state of turbulent motion.

The most important requirement of the theory is to give an explanation of the rotational motion of galaxies. In a qualitative form, the vortex theory has been propounded by Gamow (1952, 1954), by von Weizsäcker (1951), and by others. More recently, Ozernoy and Chernin (1967, 1968), Oort (1970), Sato, Matsuda, and Takeda (1970), Sato (1971), Tomita *et al.* 1971, and others have developed the vortex theory in detail in the framework of the contemporary description of the hot Universe. For example, Ozernoy and Chernin have coined the phrase "photon vortices" because the RD stage plasma is dominated by photons.

Thus, it is assumed in the modern version of the vortex theory that, before recombination, turbulent motion occurs. It is further assumed that the motion is subsonic, $u < b$. Remember that the velocity of sound during the early stage is $b \approx c/\sqrt{3} \approx 0.58c$, while close to recombination, $b \approx 0.58c(1 + 14\,\Omega_0)^{-1/2}$, i.e., $b \approx 0.15c$ for $\Omega_0 = 1$. Subsonic turbulence corresponds to incompressible fluid motion: separate volumes of fluid flow around one another. The pressure differences are then of order ρu^2. These are sufficient to change the speed and direction of the flow. But for such pressure differences, $u \ll b$ and noticeable density differences also fail to arise:

$$\frac{\delta\rho}{\rho} \approx \frac{\delta P}{P} \approx \frac{\rho u^2}{P} \approx \frac{\rho u^2}{\rho b^2} \approx \frac{u^2}{b^2} \approx 3(1 + 14\Omega_0)\,\frac{u^2}{c^2}\,. \qquad (15.9.1)$$

The picture changes sharply at recombination, after which the photons and the neutral hydrogen move independently.

The mean free path of a photon becomes enormous after recombination. Photons from different volumes intermingle; their trajectories cross. But here, in a chapter devoted to the subject of galaxy formation and related topics, the photons no longer interest us. We therefore proceed to a consideration of matter. Immediately after recombination the neutral hydrogen inherits the distribution of velocities from the preceding stage. Now the sound speed in neutral hydrogen at a temperature of about 4000 K is small: $b_{\rm H} = 0.75 \times 10^6$ cm s^{-1}. The velocity of gas motion after recombination is thus many times greater than the sound velocity, so the turbulence becomes supersonic. In other words, the pressure of the neutral hydrogen is small. It is less than the pressure of the RD plasma by just the ratio of the density of photons to that of baryons, which is about 10^9. One can therefore neglect the influence of the neutral hydrogen pressure on the gas motion after recombination. The subsequent motion takes place in accordance with the inertia of the gas as modified by the action of gravity. Now, the laws of hydrodynamics in the limiting case of zero pressure do not differ from the laws of motion of an aggregate of material points independent of

one another. In very scientific language, this is the same as saying that Newton's equations of motion are the equations for the characteristics of a system of equations in the partial derivatives of hydrodynamics for $P = 0$; (see Zel'dovich and Myshkis 1973).

Immediately after recombination the motion remains incompressible. Moreover, in homogeneous (with constant density) matter with $P = 0$, the peculiar (random) velocity of each particle decays with the cosmological expansion of $u \sim a^{-1} \sim t^{-2/3} \sim (1 + z)$, where a is the scale factor.

For the process of interest to us, the formation of isolated bodies, what plays a decisive role is that the motion *does not remain incompressible.* Initially ($t = t_r$) when the motion is free, $\nabla \cdot \boldsymbol{u} = 0$ so $\partial\rho/\partial t = d\rho/dt = 0$. However, $\partial(\nabla \cdot \boldsymbol{u})/\partial t \neq 0$, which means that the second derivative, $\partial^2\rho/\partial t^2 \neq 0$, depends on the coordinates. In the course of time, therefore, inhomogeneity arises in the density distribution. At the beginning this inhomogeneity grows in proportion to $(t - t_r)^2$. From dimensional considerations it is clear that in this period

$$\frac{\delta\rho}{\rho} \sim \frac{u^2}{l^2}(t - t_r)^2 . \tag{15.9.2}$$

In this sense the effect is nonlinear: the square of the perturbation velocity (the initial vortex velocity) enters. For the subsequent evolution, it is essential that without gravitational instability the scale increases and velocity decreases as the expansion proceeds. Therefore, the growth of the density fluctuations slows and does not continue in the unbounded way indicated by the factor $(t - t_r)^2$. In a mathematical appendix to this section, the formal aspect of the matter will be examined in detail. Here we confine ourselves to a qualitative description.

Two limiting cases are possible. The first is that of weak turbulence. Here the density fluctuations stop at a magnitude $\delta\rho/\rho \ll 1$, which requires that $ut_r/l < 1$. In other words, the turbulence itself turns out to be insufficient to cause the isolation of galaxies, galaxy clusters, and so forth. In reality, the density fluctuations—even small ones—also give rise to a corresponding gravitational field. Thus, these fluctuations represent the first stage in the subsequent growth of potential-type perturbations, a growth enhanced by gravitational instability. This limiting case therefore passes directly into an adiabatic perturbation scenario. The growth of density perturbations due to gravitational instability is common to adiabatic and to entropy perturbations. In the vortex theory the initial perturbation spectrum is cut off at small wavelengths by photon viscosity. This cutoff makes the weak turbulence theory more analogous to the adiabatic theory than to the entropy theory. The difference is that the initial velocity field contains vortex components. The circulation of the velocity is constant and nonvanishing. But the potential-type (irrotational) velocities grow in the course of time as $t^{1/3}$. Meanwhile, the vortex component of the velocity

decreases inversely in proportion to the scale, i.e., as $l^{-2/3}$, as the expansion proceeds. Therefore, the *relative* vortex velocity turns out small. In short, the theory of weak turbulence is devoid of just that main attractive feature for whose sake the turbulence is introduced. This theory does not explain the observed rotation of galaxies (Peebles 1971*b*).

Only the second limiting case, the theory of strong turbulence, can aspire to the role of a vortex theory of galaxy formation. In this theory, $ut_r/l \gtrsim 1$. Use the condition $t_r^{-1} = 3H_r/2$, where H_r is the Hubble parameter at recombination, and remember that the product lH_r is u_H, the Hubble velocity for the scale l. The strong-turbulence condition is therefore that the turbulent velocities are greater than the Hubble velocity over the turbulence scale l.

In the roughest approximation, one can neglect the expansion and the damping of the turbulent motion connected with it. After recombination, particles move in accordance with the law of inertia. As time passes, the trajectories of nearby particles intersect and infinite-density layers arise. In essence, the "pancake" situation (see § 14.2) is repeated.

The qualitative difference is that the motion is of the vortex type. Rigid-body rotation of a whole pancake is possible; so also is the sliding of some layers of the pancake with respect to others. The quantitative difference between the vortex theory and the theory of adiabatic perturbations is that galaxy formation is predicted to occur earlier in the former theory. The prediction is for this to occur when $z > 100$ (at $z = 100$, therefore, it is "now or never"). For such early pancake formation, the density is large. On the average, it is $10^{-23}\,\Omega_0$ g cm^{-3} when $z = 100$, i.e., it is not less than 10^{-24} g cm^{-3}. The density of the matter compressed into pancakes is evidently many times greater still. Furthermore, when z is large and the radiation density is large, the matter must be a strong heat emitter since it is in the form of a gas heated by a shock wave. The energy transferred by the Compton scattering of electrons must be added to the energy radiated by bremsstrahlung, emitted in electron-nucleon collisions. Compton scattering is important because the radiation density is large. Indeed, when $z = 100$, $\epsilon_r = 4 \times 10^{-13} z^4$ ergs cm^{-3} $= 4 \times 10^{-5}$ ergs cm^{-3}. The time for the loss of energy by an electron is meanwhile $(m_e c)/(2\sigma_T) = 5 \times 10^{11}$ s, which is 600 times less than the cosmological time $t \approx 3 \times 10^{14}$ s. The rapid cooling means that the developing lumps lose their heat and become flat and dense. In short, the large density of isolated bodies is the main feature and the characteristic feature of the vortex theory, and perhaps also a difficulty associated with it (see in particular Peebles 1971*b*).

In a series of investigations, Ozernoy and Chernin (1967, 1968), Ozernoy and Chibisov (1970), Ozernoy (1971), Chernin (1971, 1972), Kurskov and Ozernoy (1974*a*, *b*, *c*), and Tomita (1973) have developed the vortex theory in great quantitative detail. These authors believe that the optimum choice for the characteristic velocity is about $(0.05\text{–}0.02)c$. This specification of

one numerical parameter, together with general principles, essentially determines the remaining parameters according to the authors of these theories. See § 15.10 for more about this.

Following these authors, let us require that the indicated characteristic velocity occur at the scale size $L = ut_r$ at recombination. More precisely, it must occur at the instant of equality of the density of matter and the density of the radiation, which occurs when $z = z_{eq} = 2 \times 10^4 \Omega_0$.

On scales $R > L$, the vortex velocities are considered small. On these large scales, moreover, the initial velocities do not manage to change. Though in principle any assumptions are possible, a falling spectrum is the most economical assumption.

On scales $R < L$, the spectrum must be of the Kolmogorov (1941) type, for which $u(R)/u(L) = (R/L)^{1/3}$. As is well known, such a form follows because large-scale motion decays as a result of turbulent viscosity. The energy associated with the large-scale ($\sim L$) motion does not transform directly into heat; rather, it first transforms into the energy of smaller scale motion. A flux of energy thus occurs with respect to size scales. Now, in order of magnitude, the turbulent viscosity is $\nu_t = Ru(R)$. Here $u(R)$ is the difference between the velocities at points a distance R apart. Thus the rate of energy dissipation is

$$q = \rho\nu_t \left(\frac{\partial u_i}{\partial x_k}\right)^2 \approx \rho u R \left(\frac{u}{R}\right)^2 = \rho \frac{[u(R)]^3}{R}. \qquad (15.9.3)$$

From the condition that q not depend on R (all of the dissipated energy passes through all of the scale sizes), the law $u(R) \sim R^{1/3}$ follows. The minimum size R_0 down to which the law is valid is defined as the scale size at which interference caused by the true (molecular or photon but not turbulent) viscosity becomes important. From this follows the relation $R_0 = L[R_R(L)]^{-3/4}$.

This R_0 is less than the Silk scale $R_S = L[R_R(L)]^{-1/2}$ which corresponds to decay after the cosmological time t. The turbulent undercurrent of energy is related to the smaller time. Here R_R is the Reynolds number. We are examining R_R over various scale lengths, for which $R_R(R) = u(R)R\nu^{-1}$. In the formula presented in the text, the value of R_R associated with the maximum scale L enters. Thus, the whole spectrum of turbulent motion is known.

Let us consider first a very oversimplified case. We shall assume that the condition for radiation domination is fulfilled before recombination and that recombination takes place instantaneously. These simplifications are required so that the isolation process in neutral hydrogen begins with a Kolmogorov spectrum of vortex velocities. The largest scale size L with which an energy flux is associated has a peculiar vortex velocity just equal to the *Hubble velocity* $u_L = HL$ on the given scale. This means that the time required for a density contrast $\delta\rho/\rho \sim 1$ to arise is just of the order

of the cosmological time. On smaller scales in the Kolmogorov spectrum, the characteristic time τ is less! As a matter of fact, $\tau \sim R/u$. Since $u \sim R^{1/3}$, $\tau \sim R^{2/3}$ follows. Remembering that $\tau = t$ for $R = L$, we then obtain

$$\tau = t(R/L)^{2/3} .$$

This means that on some smallest scale l_s shock-wave-induced, strong compression and subsequent cooling of the compressed matter takes place very rapidly—in a time substantially less than the cosmological time.

Are the emerging "micropancakes" gravitationally bound? Do they transform after very little time into stars? Or does the motion along the other axes and, in particular, rotation, give rise to dispersal of these comparatively low mass but dense formations? Is the strong turbulence theory resembling entropy fluctuations (particularly the Gott and Rees 1975 formulations)? At the present time, no clear answer to these questions are forthcoming. It is important to emphasize that motion on a large scale—up to the maximum associated with an energy flux—continues normally, independent of the fate of small structures.

As time passes, perturbations satisfying $\delta\rho/\rho \sim 1$ over large scales arise by means of clusters of smaller bodies. If smaller scale processes leave the matter in a gaseous state up to this instant, then collisions of the gaseous masses and transformation—upon collision—of the kinetic energy into luminous energy continue. In this case, clusters of galaxies (corresponding to the maximum scale) also turn out to be quite dense. This corresponds to formation when $z \sim 100$.

One might expect a smaller density for clusters if galaxies and stars manage to form over small scales. The dynamics of clusters, i.e., the dynamics of large scales, is in this case the dynamics of a collisionless gas, the molecules of which are small structures—galaxies. Indeed, the possibility that trajectories cross without the particles then sticking together leads to the result that the energy of the chaotic motion of the particles does not disappear. In agreement with the virial theorem, this means that the chaotic motion limits the average density of clusters. Thus, the vortex theory in this variant leads to a picture of formation of the largest structural units—clusters and superclusters—by means of the "clustering" of smaller structures.

Ozernoy and Chibisov (1970) make the vortex theory more precise by taking account of two points. First, for $\Omega_0 > 0.07$ (corresponding to $H_0 = 75$ km s^{-1} Mpc^{-1}), equality of the matter and radiation densities, $\rho_r = \rho_m$, is achieved when $z = z_{eq} = 2 \times 10^4 \Omega_0$. Second, because the recombination is not instantaneous, additional damping takes place. During the period $z_{eq} > z > z_r$, or $t_{eq} < t < t_r$, a decrease in all velocities in proportion to the decrease in z occurs. On the other hand, recall that in the RD period the velocities of large scale motions remain constant. Neither photon nor turbulent viscosity plays a role in these large-scale motions.

The velocities also decrease during recombination. It is essential that, in the course of the change in the fraction of free electrons from 1 to 10^{-4}, the transparency change by a factor of 10^4. This of course greatly affects the interaction of the radiation. For such a perturbation wavelength, then, there is a particular instant and a particular ionization level for which the given perturbation decays most rapidly. Given these considerations, Chibisov (1972*b*) obtains the expression $u_{\text{after}}/u_{\text{before}} = \exp\,[-(M/M_*)]$ for the decay of small vortex perturbations. Here $M_* = 10^{11}\,M_\odot$.

An account of the factors enumerated thus permits Ozernoy and Chibisov to improve the agreement between the deductions of the vortex theory and observations. They notice that the theory gives (*a*) a reasonable relation between the amount of matter in rapidly rotating (spiral) and weakly rotating or nonrotating (elliptical and spherical) galaxies; (*b*) a reasonable value for the average density of galaxies and for their angular momentum; and (*c*) a reasonable distribution of galaxy clusters with respect to mass and density.

It is possible that a certain excess of precision is ascribed to the studies cited. Certainly the statistical character of the problem has been insufficiently treated. Indeed, as Doroshkevich and Shandarin (1974) have shown, the transition from the Fourier spectrum of the perturbations to the distribution with respect to mass is by no means straightforward. There are large dispersion and numerous dimensionless factors to be concerned about. The results depend not only on a single parameter, the velocity of motion on the maximum scale where there is an energy flux, but also on the behavior of the original spectrum over large scales.

However, we feel that the difficult process of making the theory more precise and the confrontation of the theory with observations of galaxies and clusters of galaxies are not the only (nor, perhaps, even the main) criteria demonstrating the correctness of the theory. It is necessary to draw together observations and theoretical considerations of another type—those based on distortions of the RR and on the behavior of a relativistic cosmological singularity.

15.10 A Comparison of the Evolutionary Theories of Galaxy Formation

In the preceding sections we have become familiar with the principal theoretical assertions concerning the structure and formation of galaxies.

A majority of these theoretical assertions are consequences of the initial conditions specified in some way or other at a more or less early time. Such a formulation of the problem results from the lack of a fundamental theory for choosing initial conditions.

At the same time several theories exist—and they were set forth above—which are each "correct" in the narrow sense that the consequences are

obtained from the assumed initial conditions. But this does not interest us so much; rather, what does interest us is the actual sequence of events that takes place in the past evolution of the Universe. In nature, there is most likely realized one certain type of initial condition, and one certain type of evolution. At the very least one type is the most important and fundamental. It seems most unlikely that two or more different types of perturbations are equally important in the past. Of course, this is a matter of personal taste.

One can in principle seek an answer to this question by comparing the calculations of the evolution with observational data. It is very important, however, to make use not only of the data concerning the mass, density, and rotation of galaxies and clusters of galaxies but also of the data concerning the RR and nucleosynthesis, i.e., processes taking place near the singularity.

One cannot construct a theory of galaxy formation while ignoring or underestimating the general knowledge of the evolution of the Universe that has accumulated up to now. In this section, therefore, the information discussed in the first two parts of the book will be used widely. Also used widely will be the information about galaxies, about the RR, and about the singularity, information that is the subsequent section, in subsequent chapters of Part III, and in Parts IV and V. The concomitant inconvenience to the reader is compensated by the realization that the general discussion of galaxy formation is not divorced from the accounts of the particular theories given in the preceding sections.

At the present time there is no theory that is unanimously accepted. The subsequent account without doubt bears the imprint of subjectivity, for we indeed have a preference for a certain set of views. Nevertheless, we shall, in the course of the exposition, set forth the reasons behind these preferences.

In constructing a general picture of galaxy formation, the first choice that one can make is the choice between processes involving the condensation of dilute gas and those involving the explosion of a superdense body. Ambartsumyan has been a consistent advocate of the second direction. To his great credit, Ambartsumyan pointed out the activity of galactic nuclei and the resemblance between processes in quasars and in galactic nuclei. Extrapolating these views, Ambartsumyan posed the question of whether one can assume that galaxies as a whole develop from superdense bodies, the remains of which (with their activity and energy preserved) are the nuclei observable at the present time. In the confines of evolutionary cosmology, it seems possible to describe these superdense bodies as, for example, matter in a singular state. By comparison with the expansion of the main part of the matter in the Universe (Novikov 1964*b;* Ne'eman 1965), the matter in these nuclei is held back in its expansion. Such retarded nuclei have been named "white holes." Some people have expressed the

opinion that perhaps only quasars and not galaxies are described in this manner.

What is the present contemporary attitude to such a line of reasoning? Let us illustrate the difficulties faced by such an approach. As an example, consider our Galaxy, with a mass $\sim 2 \times 10^{11} M_{\odot}$, and an angular momentum of the order of $\sim 2 \times 10^{74}$ g cm^2 s^{-1}. This corresponds to a circular velocity of ~ 300 km s^{-1} at a radius of about 10 kpc. Imagine that the Galaxy is formed in the past from an initially superdense body with a radius of, let us say, 1 pc. The angular momentum of the Galaxy corresponds then to a linear velocity at the equator of such a nucleus that is greater than the speed of light—a glaring contradiction. The contradiction is moreover exacerbated if one realizes that the body as a whole is located inside its gravitational radius given the mass and angular momentum of the Galaxy.

With respect to quasars, this objection is invalid because the angular momentum of quasars is unknown (although from the very fact of the compactness of the quasars it apparently follows that their angular momentum is also small). It has been shown recently, however, that the whole idea of exploding white holes encounters serious difficulty (Zel'dovich, Novikov, and Starobinsky 1974). The quantum-mechanical process of the creation of particles near the singularity turns out to lead inevitably either to the early explosion of the white hole or to no explosion of it at all.

We emphasize that the aforementioned critical remarks do not concern ideas about the present-day activity of galactic nuclei and of quasars. They do, however, limit the notions of a possible source of the energy fueling this activity.

With respect to galaxy formation, the point of view of condensation from a dilute gas is dominant today. If one accepts this point of view, then the question remains concerning the choice of the type of initial perturbations of the homogeneous and isotropic Friedmann cosmological model. That is, the choice remains between the theories of galaxy formation set forth in the preceding sections.

Let us begin with a discussion of the behavior of perturbations near the singularity. Adiabatic and entropy perturbations are compatible with small metric perturbations, i.e., with a quasi-isotropic solution. Near the singularity, such a solution is not distinguished by its local properties from a strictly homogeneous and isotropic solution. Consequently, the use of the quasi-isotropic solution does not change the known results concerning nucleosynthesis (see § 7.5) that are calculated in the context of a strictly homogeneous theory. Note too that these results, namely, an abundance by mass for ^{4}He of 0.25–0.27, and for D less than 10^{-4}, are apparently supported by observations (see § 7.6).

The vortex theory requires a substantially anisotropic singularity (see § 12.4, § 12.6), so vortex perturbations necessarily require changes in the expansion law near the singularity. This leads (see § 21.5) to primordial

abundances that differ sharply from that cited above and that probably are not in accord with observations (either there is more than 30% ^{4}He or there is 2–3% D and 97–98% H). But something still more important occurs. Imagine that the beginning of the expansion is actually non Friedmann-like and that only in the course of time does the solution approach Friedmann behavior. In accordance with the vortex theory, all of the deviations in the metric from the Friedmann form, except those of the vortex variety, are sharply suppressed. One can then ask: Why are the initial conditions at the singularity of just such a form that they lead in the distant future to such a special situation? It is much more plausible to assume that all types of perturbations are equally represented (the uniform-distribution hypothesis) since a solution with a non–Friedmann-like beginning approaches a Friedmann form. But then, as we saw in chapter 12, vortex perturbations rapidly decay while entropy perturbations remain constant and adiabatic perturbations grow. The evolution subsequently follows the tracks laid out by the theory of perturbations.

Finally, it is important to take account of some additional quantum phenomena that occur near the singularity. The creation of particles at an anisotropic singularity leads, in the simplest case, to isotropization, thereby switching the cosmological model to the quasi-isotropic solution with, perhaps, variable entropy (see chap. 24). With respect to concrete results relevant to cosmological models with matter in vortex motion, much research will have to be done. It is quite likely that models with vorticity will, as a result of internal contradictions, disappear entirely when account is taken of quantum phenomena. See, in this respect, Lukash, Novikov, and Starobinsky (1975).

Consider now the viewpoint of an astronomer who does not take account of the arguments relating to the period from the singularity to the time of nucleosynthesis. We shall analyze the later period beginning with the termination of electron-positron annihilation and extending up to the present time. This is the period when $10^8 \gtrsim z \gtrsim 0$ and includes the epochs of the RD plasma; the recombination of helium and hydrogen; the growth of perturbations in the neutral gas; the formation of galaxies, quasars, and so forth; and the secondary ionization of the neutral gas that is not a part of gravitationally bound objects. We shall begin with a comparison of the deductions of theory with the RR observations.

A distinctive feature of the vortex theory is the assumption of a large perturbation amplitude of a large velocity for the vortex motions that are superposed on the overall Hubble expansion. Another feature, an unavoidable one, is that vortex motion does not give rise to density perturbations immediately after recombination.

In the first variant of the theory it was assumed that the velocity of the vortex motion at recombination is of the order of 0.1–$0.4c$. The corresponding RR temperature fluctuations due to the Doppler effect are

evidently $\Delta T/T \sim 0.1$ to 0.1. But nothing like this is observed; actually, $\Delta T/T \sim 3 \times 10^{-4}$. To explain this, the authors of the vortex theory assumed that the temperature fluctuations are diminished by Compton scattering. The smoothing of RR fluctuations by subsequent scattering is treated by Ozernoy and Chernomordik (1976). It is necessary for the fluctuations to be reduced by a factor of at least 300, thereby requiring not less than six scatterings or an optical depth $\tau \geq 6$. In addition, the large initial velocity actually leads to the collision of gas clouds and the formation of shock waves very shortly after recombination.

In the first variant of the vortex theory considered above, it was assumed that galaxies form quite early, for $z \sim 130$. The early collision of clouds and the early formation of gravitationally bound objects at first glance seems compatible with the repeated RR scattering that is necessary in order for the large vortex velocities of the RD plasma to agree with the observable RR isotropy. In reality, however, the early formation of shock waves and the early re-ionization of the gas give rise to great difficulties. At early times there is also high radiation density and a rapid cooling of the electrons by inverse Compton scattering from photons, the scattering involving an increased photon frequency (on the average) and an electron energy loss. As the electrons cool, therefore, the plasma density correspondingly grows since the cold plasma is surrounded by a hot, still uncooled plasma, and the pressure is conserved in the first approximation. So, for example, with $\Omega_0 = 0.5$, when $z = 130$ the average density is 3 cm^{-3} while in the shock wave the density grows to 12 cm^{-3}. For a speed of 10^9 cm s^{-1} (the initial speed is $u \approx 0.3c = 10^{10}$ cm s^{-1}, and it falls during the expansion in proportion to $\sim z$), a temperature of the order of 10^9 K is achieved. Upon the fall of the temperature to 10^5 K, the density increases 10^4 times, achieving the value 10^5 cm^{-3}. At such a density, recombination proceeds quite rapidly and it is impossible for the gas to be maintained in a completely ionized state for an extended period.

Later another variant of the vortex theory was proposed, the so-called slow-motion variant. This variant is distinguished by the comparatively low initial vortex velocity of $u/c \sim 0.03$.

In the postrecombination neutral gas the vortex velocity decays inversely in proportion to linear scales. Recall that any velocity superposed on the overall expansion decays by such a law (though the velocity decreases, the angular momentum is conserved!) if the gravitational field of the density perturbation does not cause this velocity to increase. Suppose that $ul/\lambda < 1$ at recombination. This condition is equivalent to the condition that the vorticity (the magnitude of $\nabla \times \boldsymbol{u}$ is of the order of the ratio u/λ) is less than the Hubble parameter $H \sim t^{-1}$. Then, to an approximation that does not take account of the gravitational field of the perturbations, we find that the shocks never manage to develop (see the Appendix to

§ 15.9). So, for $t \to \infty$, we obtain

$$\left.\frac{\delta\rho}{\rho}\right|_{\infty} \approx \left(\frac{u_0 t_0}{\lambda_0}\right)^2 \ll 1 . \qquad (15.10.1)$$

The general overall picture of the slow-motion variant is then as follows. An initial vortex perturbation decays, leaving a small density perturbation. Following this, the density perturbation gives rise to potential motion that strengthens the density perturbation via the gravitational-instability mechanism (see the Appendix to § 15.9). Upon selecting an initial perturbation, we will obtain a certain amplitude for the growing potential-type perturbation. In the slow-motion variant this choice is presumably made in such a way that the formation of gravitationally bound objects is timed to occur when $z \sim 20$–30.

But at the same time there again remains a contradiction with the observed RR isotropy. The choice of $u/c \sim 0.03$ initially would give $\Delta T/T \sim 0.03$, which is 100 times greater than what is observed! In the slow-motion variant the necessity for a smoothing-out of the initial radiation anisotropy has therefore not vanished. If, however, the formation of galaxies and the secondary ionization of the gas takes place when $z \approx 30$ or later, then it is impossible for the anisotropy to be smoothed out. This is because the maximum optical depth is, in reality,

$$\tau_{\max} \approx 0.06(\Omega_0 h_0^2)^{1/2} z^{3/2} , \qquad (15.10.2)$$

where $h_0 = H_0/(100\ \mathrm{km\ s^{-1}\ Mpc^{-1}})$. With $\Omega_0 = \frac{1}{2}$, $z = 30$, and $h_0 = \frac{1}{2}$, we obtain $\tau_{\max} \approx 3.5$. Therefore, $e^{\tau} = 32$, while 100 is necessary in order for the theory to agree with the observed RR isotropy.

In short, the vortex theory is apparently not compatible with the observed RR isotropy.

Along with the question of whether the vortex theory is possible, a question raised by a comparison of the theory with the RR observations, one must also consider the question of the attractiveness of the theory. Is there not an element of arbitrariness in the theory? What are its other consequences besides the predictions concerning the mass and rotation of galaxies and the RR anisotropy? It is necessary to compare the vortex theory with the theory of adiabatic perturbations.

To explain the present-day structure of the Universe in terms of the theory of adiabatic perturbations, one must specify the metric perturbations in such a way that they have an amplitude of the order of 10^{-3} to 10^{-4} on a scale corresponding to a mass $M \sim 10^{13}$–$10^{14}\ M_{\odot}$. One can continue the perturbation spectrum smoothly with the same amplitude (with respect to the metric) to higher and lower masses. Perturbations with longer wavelengths, corresponding to masses $M \gg 10^{14}\ M_{\odot}$, grow more slowly and result in $\delta\rho/\rho \ll 1$ today (given the same metric perturbation amplitude). They also produce RR perturbations of the order of 10^{-4}, which

does not contradict observations (see the following chapter). Perturbations corresponding to masses $M \gtrsim 10^{14} M_\odot$ decay and do not affect the creation of galaxies (see § 11.2). If the amplitude of these perturbations is also of the order of 10^{-3} to 10^{-4}, then upon their decay no observable effects remain.

Thus, in the adiabatic theory, one can choose an initial spectrum (without any preferred scale) by giving one characteristic quantity—the amplitude—and by leaving the selection of the scale of clusters of galaxies to the laws of nature. It is instructive to compare these properties of the theory of adiabatic perturbations with the properties of the vortex theory.

The authors of the vortex theory maintain that this theory contains only one initial parameter—the characteristic velocity u or the dimensionless ratio $\beta = u/b \approx 2u/c$ of the vortex velocity to the velocity of sound, b, at the moment when the radiation density and the matter density are equal. It is assumed that this velocity applies to a scale related to the velocity by $R \sim ut$.

The most important fact is that in the vortex theory it is necessary to take the initial turbulence spectrum (at the beginning, during the RD period, when $z \gtrsim 10^8$) with a preferred scale R and with a sharp maximum of $\beta^2(r)$ at $r = R$. The necessary value of β mentioned above is also this maximum, and we denote it by β_m. It is necessary to take the initial spectrum with $\beta(r) < \beta_m$ for $r \ll R$ in order that the initial, small-scale turbulence not distort the RR upon decay. But it is also necessary to limit the perturbation spectrum of the vortex theory on the side of larger scales $r > R$. This limitation follows from the absence of noticeable angular fluctuations in the RR.

We thus see that the vortex theory requires a special choice of initial spectrum and, in this sense, compares unfavorably with the adiabatic theory.

From this account it should be apparent that the present authors prefer the adiabatic theory of galaxy formation.

We have not discussed the entropy theory of galaxy formation here (see § 15.1). In recent years new observations and theoretical analysis have made the entropy theory more and more attractive. The final choice among these theories lies in the future.

15.11 Observational Data concerning the Properties of Galaxies and Galaxy Clusters; the Average Density of Matter in the Universe

In this section we shall present in very brief and schematic form a summary of certain observational data concerning the basic properties of galaxies and galaxy clusters. A recent and comprehensive survey of extragalactic astronomy is the ninth volume of the series *Stars and Stellar Sys-*

We thank Ya. Einasto and E. Saar for help in preparing § 15.11.

tems (Sandage, Sandage, and Kristian 1975). As a review of extragalactic astronomy, we recommend, for example, the book by Vorontsov-Vel'yaminov (1972), the book edited by Pikel'ner (1976), the earlier review by Abell (1962), and also the popular book by Agekyan (1970). Moreover, Field (1973*b*) has reviewed data relating to intergalactic gas, while observational data concerning the average density of matter in the Universe is presented in Peebles's (1971*a*) book.

We must forewarn the reader that many of the parameters have not been determined with great certainty and thus the values vary from author to author. Moreover, the error estimates are very subjective.

An important characteristic property of galaxies that is relatively simply established from observations is the luminosity function—the number of galaxies with a given absolute magnitude, or, more precisely, with a magnitude between M and $\mathsf{M} + d\mathsf{M}$. Peebles (1971*a*) cites the following approximate formula for the average number of galaxies per unit volume brighter than absolute magnitude M. This formula is

$$\begin{aligned} n(<\mathsf{M}) &= A \cdot 10^{\alpha \mathsf{M}}, \quad \alpha = 0.75, \quad \mathsf{M} < \mathsf{M}^*; \\ n(<\mathsf{M}) &= B \cdot 10^{\beta \mathsf{M}}, \quad \beta = 0.25, \quad \mathsf{M} > \mathsf{M}^*. \end{aligned} \tag{15.11.1}$$

For $\mathsf{M} = \mathsf{M}^*$, it is evident that $A \cdot 10^{\alpha \mathsf{M}} = B \cdot 10^{\beta \mathsf{M}^*}$. The absolute visual magnitude M^* is

$$\mathsf{M}^* = -19.5 + 5 \log h_0, \tag{15.11.2}$$

where $H_0 = 100 h_0 \text{ km s}^{-1} \text{ Mpc}^{-1}$.

To theoreticians, parameters related to the masses of galaxies have great significance. The most reliable path to a determination of the masses consists of measurements of the line-of-sight velocities of rotation of galaxies. Then, given the known linear size of a galaxy, one can compute the mass according to Newton's law. But only irregular and spiral galaxies rotate noticeably. The masses of elliptical galaxies can be determined via the dispersion in the line-of-sight velocities of the stars. The mass function for galaxies of various types is presented in Figure 52; it was obtained by Holmberg (1969). The masses of the most massive galaxies are about $10^{12} M_\odot$ or even more. These include the E giants and also Sb's (see Vorontsov-Vel'yaminov 1972). Galaxies with a mass of about $10^{11} M_\odot$ make the greatest contribution to the total mass. Dwarf spheroidal galaxies with $M \approx 10^5 M_\odot$ are galaxies with the minimum mass.

Table 12 (Holmberg 1969) shows the contribution to the total mass of galaxies with various masses.

The mass-to-luminosity ratio M/L (with M and L in solar units) is about 5 for irregular (type I) galaxies, about 7–20 for spiral (type S) galaxies, and about 40 for elliptical (type E) galaxies.

Radio observations indicate that rotation curves of most spiral galaxies are very flat: $V(R) = \text{const.}$ up to large galactocentric distances $R \sim 50$ kpc

(Roberts and Whitehurst 1975; Huchtmeier 1975; Salpeter 1977). Hence the cumulative (inner) mass $M(R) \propto V^2(R)R \propto R$. On the other hand, the spatial luminosity density in disks of galaxies is well represented by an exponential model $I(a) = I_0 \exp(-a/a_0)$ (Freeman 1970). Hence the cumulative luminosity

$$L(R) \propto \int_0^R I(a)a^2 da$$

converges rapidly to the total luminosity of the galaxy L_{tot}. The cumulative mass-to-luminosity ratio $M(R)/L(R)$ has, inside of the optically visible boundaries of spiral galaxies, a value between 3 and 10 (in solar units), depending on the prevailing stellar population (for young populations and old metal-poor populations M/L has a low value; for old metal-rich populations, a value 10 or even higher). See also Faber and Jackson (1976). About the possibility that outside the optically visible regions the cumulative mass-to-luminosity ratio grows proportionally to R, see the end of this section.

The diameters of the brightest galaxies are about 3×10^4 pc. Dwarf galaxies have much smaller dimensions. The angular momentum of galaxies

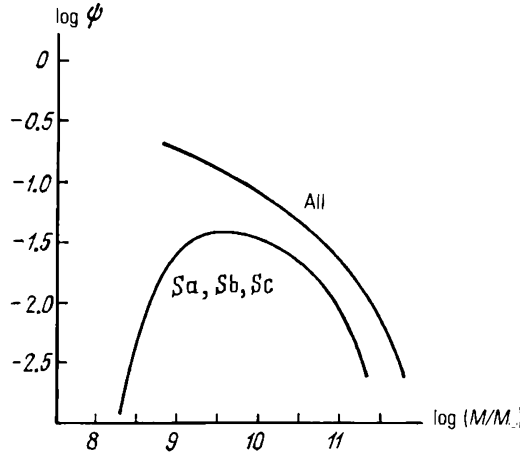


FIG. 52.—The mass function for galaxies of various types

TABLE 12

PERCENTAGE OF THE TOTAL MASS CONTRIBUTED BY GALAXIES WITH VARIOUS MASSES

	$M_\odot/M$			
	$<10^9$	10^9–10^{10}	10^{10}–10^{11}	$>10^{11}$
Percent contribution...	3	11	40	46

is known poorly. According to one recent analysis of the observations (Nordsiek 1973), the specific angular momentum of spiral galaxies is related to the mass by

$$\log \mathfrak{M} = -0.4 + 0.67 \log M . \qquad (15.11.3)$$

Here $\mathfrak{M}$ is the angular momentum in units of $10^{12} M_\odot$ kpc km s^{-1} = 6.2×10^{71} g cm^2 s^{-1} and M is the mass in units of $10^{10} M_\odot$. The distribution of galaxies with respect to type is given in Table 13 (van den Bergh 1960), which includes data for 935 galaxies brighter than an apparent magnitude m of 13. Of course, the detailed table entries, based upon a count up to a definite magnitude, do not reflect directly the distribution of galaxies with respect to type per unit volume of space. It has been shown that the distribution of galaxies with respect to type depends very strongly on the type of galaxy cluster (see more about this later).

Zwicky, Herzog, and Wild (1968) have published a catalog of several thousand galaxy clusters. Earlier, Abell (1958) divided clusters into regular and irregular ones. Besides this important separation of clusters into two groups, other types of cluster classification exist. These other classifications are, for example, based on richness (number of members), the presence of peculiar galaxies, etc.

Properties of galaxies are closely related to the properties of their respective systems. In rich clusters of galaxies, most galaxies are of the type E or S0, which can be explained by ram pressure of the intracluster gas (van den Bergh 1977; Komberg 1976). Spiral galaxies in clusters contain less gas than similar field spirals do.

A close link between the properties of galaxies and respective systems is observed also in poor systems. Galaxies in groups are segregated according to their morphological types (Einasto, Saar, Kaasik, and Traat 1974): ellipticals tend to populate inner regions; spirals and irregulars, the outer ones, as is the case in clusters. The velocity dispersion of stars in the main galaxies of systems is equal to the mean velocity dispersion of the galaxies. This relation holds for systems of all scales from poor groups to rich clusters (Einasto *et al.* 1976*a*). The latter relation indicates that systems of galaxies form together with galaxies, since a later adjustment of dynamics takes too long.

TABLE 13

DISTRIBUTION OF GALAXIES WITH RESPECT TO TYPE

	Galaxy Type					
	E and S0	Sa	Sb	Sc	Irr	Others
Percentage.....	22.9	7.7	27.5	27.3	2.1	12.5

Regular clusters consist of a large number (sometimes more than 10^4) of members, are spherically symmetric and have a greater concentration at the center. The bright members of these clusters are probably only type E and S0 galaxies. According to Vorontsov-Vel'yaminov, spiral galaxies can only be projected upon such clusters by chance. At the center of a cluster, one or two extremely bright elliptical galaxies are often found. A typical representative of the regular clusters is the cluster in the constellation Coma Berenices (the Coma cluster). Irregular (diffuse) clusters have an irregular form; distinct condensations are often encountered in them. These clusters consist of all types of galaxies. They can be both rich (greater than a thousand members), as is, for example, the Virgo cluster, or very poor with respect to number of members.

The best studied clusters are the regular clusters. Thus, according to Goldsmith and Silk (1972), the size of the Coma cluster is about 4 Mpc. The total number of galaxies in the cluster (including faint galaxies) is estimated to be several tens of thousands. The dispersion of the line-of-sight velocities of the galaxies is about $\Delta V = 1000\,\mathrm{km\,s^{-1}}$. Rood (1972) emphasizes that ΔV depends strongly on the distance from the center of the cluster, falling in value from the center to the edge. At the edge, $\Delta V = 500\,\mathrm{km\,s^{-1}}$. If one considers the clusters to be stationary, then, by using the virial theorem, one can find the cluster mass as a function of the velocity dispersion and the size. This turns out to be

$$M_{\mathrm{vir}} \approx 3 \times 10^{15}\, M_{\odot}[H/(75\,\mathrm{km\,s^{-1}\,Mpc^{-1}})]^{-1}\,, \qquad (15.11.4)$$

an unexpectedly large figure. Also one can, by a count of the galaxies, determine the integrated luminosity of a cluster and then calculate the ratio M/L. According to Abell, this ratio is about

$$\frac{M}{L} = 300[H/(75\,\mathrm{km\,s^{-1}\,Mpc^{-1}})]\,, \qquad (15.11.5)$$

which is many times larger than the value of M/L for elliptical galaxies, the galaxy type with the largest M/L value. This result apparently signifies one of the following: (1) the cluster is nonstationary, even though its regular form and its increased concentration toward the center suggest a stationary state; (2) there is, in the cluster, much invisible mass between the visible parts of galaxies; or (3) there are important systematic errors in the observations and their interpretation. A short discussion of the situation is given at the end of this section.

An example of an irregular cluster is the Virgo cluster. It consists of about a thousand members. According to Abell (1965), its size is about 3 Mpc. He also estimates the mass-to-luminosity ratio for the cluster as no less than 400.

In any test of galaxy formation theories, the question of the orientation

of the axes of rotation of the galaxies in clusters has great significance. There are no reliable data concerning this, so we confine ourselves to calling attention to work by Brown (1968), Reinhardt (1972), and Reinhardt and Roberts (1972).

The question has long been discussed of whether clusters of clusters exist. Even now the arguments have not ceased. Zwicky (1962) believes that clusters of clusters do not exist. Vorontsov-Vel'yaminov holds to the same opinion. Abell, on the other hand, adopts the opposite viewpoint. G. de Vaucouleurs (1971) demonstrated that the majority of galaxies are in systems of richness and scale varying from pairs to groups to clusters and superclusters. Apparently, however, all agree that significant density inhomogeneities on scales tens of times larger than the sizes of the largest clusters do not exist. Indirect considerations provide the strongest arguments supporting this view. These considerations are based on observations of the RR isotropy, a subject we have mentioned above and will consider again in the next chapter. In agreement with Zwicky, the diameters of the largest clusters are of the order of 8 Mpc, and the clusters are, on the average, distributed homogeneously. The size of the largest inhomogeneity in the Universe is of the order of 40 Mpc.

Slowly but steadily accumulating data on the radial velocities of galaxies have enabled a start to be made in reconstructing the three-dimensional structure of the Universe. The first studies discovered that the Coma cluster of galaxies extends to $\sim$100 Mpc (Einasto *et al.* 1975*b;* Chincarini and Rood 1975) and that its shape is flat rather than spherical (Tifft and Gregory 1976; Gregory and Tifft 1976). The spatial distribution of nearby galaxies suggests a cell structure; there are regions almost void of galaxies, the latter being concentrated at cell walls (superclusters [Jôeveer *et al.* 1977]). The mean cell diameter is $\sim$100 Mpc, coinciding with the characteristic distance from the galaxy correlation studies (Davis, Groth, and Peebles 1977; Seldner *et al.* 1977).

As we already noted, a huge number of small clusters, groups, and multiple galaxies exist in addition to the large clusters. Practically all of the groups of galaxies studied turn out to be unstable if the virial theorem is applied to them. Their kinetic energy, calculated on the basis of a measurement of the dispersion of the line-of-sight velocities of the galaxies, greatly exceeds the potential energy, calculated on the basis of the masses of the visible galaxies. Very impressive examples of unstable groups are presented in the review by M. Burbidge and Sargent (1968).

Thus, accounting only for the material contained in galaxies, all systems, from multiple galaxies to clusters and to the possible superclusters, turn out to be unstable. This is illustrated in Table 14, which was constructed by Karachentsev (1968).

The question of whether systems of galaxies are stable is closely connected

with the question of the determination of the average density of matter in the Universe. If one accounts only for the easily observed forms of matter in galaxies, then, as we have seen, systems of galaxies turn out to be unstable. At the same time, the average density of matter in the Universe $\langle\rho\rangle$ turns out to be considerably less than the critical density. Oort (1958) carried out the best early determination of $\langle\rho\rangle$, taking account only of the matter in galaxies; subsequent refinements have made no substantial changes in this estimate. Oort's determination is based on a determination of the average luminosity per unit volume in the Universe by means of galaxy counts and by the use of the average value of the ratio M/L for galaxies. If this average value is $M/L \approx 15\,[H/(75\text{ km s}^{-1}\text{ Mpc}^{-1})]$, then $\langle\rho\rangle \approx 2\times10^{-31}\,[H/(75\text{ km s}^{-1}\text{ Mpc}^{-1})]^2$ g cm^{-3} and $\langle\rho\rangle/\rho_{\text{crit}} \approx 0.02$.

The existence of a paradox concerning the instability of clusters of galaxies has led to the assumption that a large mass of invisible material is present. If there is sufficient such material, then the clusters will be gravitationally bound, and the average density of mass in the Universe is closer to the critical value.

The first candidate for such "invisible mass" is intergalactic ionized gas. The question of the possible amount of such gas was discussed in detail in Part 2. Recently, the situation was summarized in a review by Field (1973*b*). His conclusions amount to the following. Rich clusters, such as Coma, are X-ray sources. If one assumes that the emission is from a gas at a temperature $T \approx 10^8$ K, then the amount of gas is too small to bind the cluster gravitationally. The assumption that in a cluster there is a significant number of colder gas clouds with a temperature $T \approx 10^5$ K is excluded by a combination of optical, ultraviolet, and soft X-ray observations. Thus, the total mass of such gas in the Coma cluster is at least several times smaller than that required to stabilize the cluster. Theoretical considerations concerning the accretion of gas located between clusters onto rich clusters, together with observational data concerning the clusters, leads to the conclusion that $\Omega \leq 0.05$ for such gas.

TABLE 14

AVERAGE CHARACTERISTICS OF SYSTEMS OF GALAXIES

Galaxy Systems	Luminosity ($10^{10}L_\odot$)	Average Number of Members	Radius R (10^4 pc)	Dispersion of the Line-of-Sight Velocities (km s^{-1})	Observable Density (10^{-25} g cm^{-3})	Virial Density (10^{-25} g cm^{-3})	Dispersal Time $R/\Delta V$ (10^8 yr)
Triplets	4.9	3	8.8	121	0.10	1.0	4.5
Groups	10.5	8	39	287	0.002	9×10^{-2}	11
Poor clusters	43	35	114	354	4×10^{-4}	2×10^{-2}	20
Rich clusters	270	220	271	827	2×10^{-4}	2×10^{-2}	18
Superclusters	1500	1200	1640	1100	4×10^{-6}	6×10^{-4}	87

Observations still cannot exclude the possibility that a significant amount of gas with a temperature between 10^5 K and 10^7 K is present in the so called de Vaucouleurs groups of galaxies. Modern data do not contradict the possibility that $\Omega \approx 1$ due to such gas. To test this, observations are necessary in the soft X-ray band. Recall that the presence or absence of this gas has no relation whatsoever to the problem of the stability of rich clusters.

Recent observations have revealed a considerable amount of neutral gas near galaxies (Winter 1975; Mathewson *et al.* 1975; Daniels 1975; Wright 1973). A well-studied example is the gas near our Galaxy.

A part of this gas forms long streams which are connected with companion galaxies: the Magellanic Stream (Mathewson *et al.* 1975) and similar streams in the northern galactic hemisphere. As demonstrated by Lynden-Bell (1976) and by Kunkel and Demers (1976), all optical companions of the Galaxy and all hydrogen streams are concentrated in a plane, which forms almost a right angle with the plane of the Galaxy. This gas may be a remnant of Galaxy formation (Verschuur 1975). Its spatial distribution in groups of galaxies and the above-mentioned properties of groups have been inferred to separate such groups in a special class—hypergalaxies (Einasto *et al.* 1976).

Quite recently, many proponents of the idea that galaxies can be surrounded by huge halos of faint stars or other dark matter have come forth. Since such mass in a halo does not noticeably influence the dynamics of the internal parts of galaxies, which is well studied, the observations of the internal parts only give a mass of these internal parts and say nothing about the masses of halos. It is possible that taking account of these halos will substantially change the estimate of the masses of galaxies in clusters and thus the masses of the clusters themselves.

To check the massive halo hypothesis and to estimate the mass and extent of the halo, Einasto, Kaasik, and Saar (1974) and Ostriker, Peebles, and Yahil (1974) studied the motions of companions of giant galaxies. It was demonstrated that the cumulative mass varies as $M(R) \propto R$ out to a distance $R = 200$–500 kpc. The total mass of the halo is an order of magnitude greater than the mass of the central galaxy. As demonstrated by Einasto *et al.* (1975*a*) and Turner (1976), the mean mass-to-luminosity ratio of systems of galaxies (galaxies with massive halos) is about 65–100 for systems with spiral main galaxies and about 200 for systems with elliptical main galaxies. See also de Vaucouleurs (1969), Arp and Bertola (1971), Kormendy and Bahcall (1973), Tarter and Silk (1974), Ozernoy (1974), and Komberg and Novikov (1975).

The total mass density of the Universe, according to new estimates, leads to $\Omega \approx 0.2$–0.6 (Einasto, Kaasik, and Saar 1974; Ostriker, Peebles, and Yahil 1974; Turner and Gott 1976).

15.12 Astrophysical Consequences of the Existence of a Heavy Neutral Lepton

If neutrinos have a nonzero rest mass (as we shall see they might), then they become an attractive candidate for bearer of the galactic missing mass. As is known, the mass-to-light ratio for galaxies as a whole is considerably higher than for average stars in the middle part of a galaxy. The galactic mass can be determined by the gravitational attraction between galaxies in a cluster of galaxies. Another way to find the mass of a galaxy is to measure the variation of the velocities of matter with an increase of distance from the galactic center. Such measurements show too that the bulk of the matter in galaxies is outside the region where luminous objects are seen. So the dynamics of galaxies are determined by some dark, invisible matter. Marx and Szalay (1972), Szalay and Marx (1976), and Cowsik and McClelland (1972) propose that neutrinos with a mass of about 10 eV could solve the problem of the missing mass. The idea was later considered by Marx (1977), Gunn *et al.* (1978), and Tremaine and Gunn (1979). They argued that neutrinos mainly cluster with galaxies forming galactic halos.

Using the condition

$$\Omega^* = (\text{the matter density in galaxies})/\rho_{\text{crit}} < 0.1\,, \qquad (15.12.1)$$

Tremaine and Gunn conclude that $\Sigma m_\nu < 0.6$ eV. From the Pauli exclusion principle it follows that the maximum neutrino density $n_{\max}$ in phase space cannot exceed $2g_\nu$, where g_ν is the number of helicity states. With a more accurate treatment of neutrino dynamics Tremaine and Gunn conclude that $n_{\max} < g_\nu$. On the other hand, if neutrinos in a galactic halo form a Maxwellian thermal gas, then it follows from the known value of $n_{\max}$ that $m_\nu > 10$ eV. The contradiction between the upper and lower limits on m_ν led Gunn and Tremaine to conclude that light neutrinos cannot solve the problem of the missing mass. Heavy leptons, however, with a mass of a few GeV can provide the needed mass. In our opinion the accuracy of the bound (15.12.1) is not great; it could be an order of magnitude weaker. In this case the conclusion of Tremaine and Gunn could be invalid, and light neutrinos might supply the missing mass.

Recently Ljubimov and his group (1980) have claimed that a thorough investigation of the electron spectrum in tritium decay proves that the neutrino (now specifically the electron antineutrino) has a rest mass of the order of 30 eV (16 eV $< m_\nu <$ 46 eV) at the 99% confidence level. Using the data of the Reines experiment, Barger, Whisnant, Cline, and

Section 15.12 is extracted from the article "Cosmology and Elementary Particles," by A. D. Dolgov and Ya. B. Zel'dovich, *Reviews of Modern Physics*, vol. 53, no. 1 (Jan. 1981). The editor (G.S.) disclaims responsibility for the accuracy of the material here.

Phillips (1980) find oscillations in an antineutrino beam inducing the reaction $\bar{\nu}_e + d \rightarrow 2n + e^+$. The difference in the squares of the neutrino masses $(m_1^2 - m_2^2)$ was estimated to be of the order of 1 eV^2, but other solutions with a larger difference are also possible. Reines, Sobel, and Pasierb (1980) claim to have detected oscillations by comparison of the charged-current reaction $\bar{\nu}_e + d \rightarrow 2n + e^+$ and the neutral-current reaction $\bar{\nu}_e + d \rightarrow p + n + \bar{\nu}_e$ induced by a beam of reactor antineutrinos. Oscillations $\nu_e \rightleftarrows \nu_\tau$ seem also to have been observed by the CHARM collaboration (see Winter 1979).

All this taken together is changing the prevailing attitude of the astrophysical establishment toward neutrinos with nonvanishing rest mass of the order of some eV or tens of eV. Earlier it was commonplace to put upper limits on the neutrino mass assuming that it would converge to zero. Now one is tempted to explain real phenomena by invoking a neutrino rest mass. Marx and Szalay, who adopted this point of view 10 years ago, thus find themselves in the position of forerunners.

The first success of such an approach is the explanation of the deficit of solar neutrinos in the Davis experiment by oscillations $\nu_e \leftrightarrow \nu_\tau \leftrightarrow \nu_\mu$, as proposed by Pontecorvo. The prediction depends on mixing angles, but a value of $\frac{1}{3}$ for the ratio of electron neutrinos to the total number of neutrinos at the Earth's orbit is reasonable.

One obvious astrophysical implication is that if the sum of neutrino masses exceeds 25 eV, the universe is closed and its expansion must change to contraction in the remote future. The age of the universe is also shortened. To a good approximation it is

$$t = 21 \times 10^9 \text{ yr}/[(H/50) + (\Sigma/40)^{1/2}] . \qquad (15.12.2)$$

Here H is the Hubble constant in kilometers per second per megaparsec (the best value is $H = 55$) and Σ is the sum of the neutrino rest masses in eV.

If all neutrinos have the same rest mass, then when $\Sigma = 90$ one attains $t = 8.5 \times 19^9$ yr, which is less than the values found by Os/Re chronometry ($11\text{–}18 \times 10^9$ yr) and by stellar evolution theory applied to old globular clusters ($14\text{–}16 \times 10^9$ yr).

A possible outcome was suggested by Zel'dovich and Sunyaev (to be published), who note that the age of the universe can be lengthened by the introduction of the cosmological constant. One gets thereby a closed but still-expanding universe. The upper limit on m_ν which is valid for models with a nonvanishing cosmological constant is about 200 eV. Accidentally we return to the first rough estimate of Gershtein and Zel'dovich (1966).

Another way of achieving $\Sigma \gg 25$ eV without revising any fundamental ideas is to assume that an unknown source of background photons existed,

so that the ratio N_ν/N_γ is smaller than in the standard model, and consequently $\Sigma \gg 25$ eV is required to close the universe. An example of such a source would be a new long lived particle decaying into photons.

As has already been pointed out by Marx and Szalay, if $m_\nu \neq 0$ the growth of perturbations leading to galaxy formation and clustering is also changed: it begins now with the growth of neutrino density perturbations.

In a recent investigation Zel'dovich, Sunyaev, Doroshkevich, and Kholopov (1980) conclude that $m_\nu \neq 0$ permits galaxies to form from smaller initial perturbations. This makes it easier to understand the negative result of all searches for background temperature fluctuations.

Neutrinos more massive than those considered by Tremaine and Gunn (1979) are good candidates for the explanation of the "hidden mass" in galaxies and clusters of galaxies.

Despite the fact that massive neutrinos could solve some cosmological problems, we would like to end this discussion with a warning: astronomy will use gratefully any neutrino rest mass found in laboratory experiments, but astronomy cannot be used to prove or even to support a definite value of or even the very existence of a nonzero mass for neutrinos or some other unknown particles.

MATHEMATICAL APPENDIX TO § 15.9

Here we consider the nonlinear regime of the vortex perturbations (postrecombination). The pressure is neglected ($P = 0$).

Let us find the function $f(t)$ characterizing the peculiar motion of matter in an expanding Universe. We shall do this in the approximation where the gravitational fields connected with the density perturbations are neglected. This disregard also distinguishes the problem from the problem considered earlier (§ 14.2) of the nonlinear theory of growing adiabatic perturbations.

Thus, we seek ($\boldsymbol{r}$ is the Eulerian coordinate of a particle, $\boldsymbol{q}$ is the Lagrangian coordinate)

$$\boldsymbol{r} = t^{2/3}\boldsymbol{q} + f(t)\psi(\boldsymbol{q}). \tag{15.9A.1}$$

The first term describes the Hubble flow, the second describes perturbations. The absolute velocity of a particle in the Newtonian picture is $d\boldsymbol{r}/dt$; with respect to this differentiation, $\boldsymbol{q}$ is a constant vector. The peculiar velocity is meanwhile the difference between the Hubble velocity of the unperturbed medium and the absolute velocity at the same point in space:

$$\boldsymbol{u} = \frac{d\boldsymbol{r}}{dt} - H\boldsymbol{r} = \left(\frac{df}{dt} - \frac{2}{3t}f\right)\psi\,. \tag{15.9A.2}$$

As is well known (see Part I), the peculiar velocity decreases inversely as the scale factor, i.e., as $t^{-2/3}$. Incorporating a numerical factor in the definition of the function ψ, we obtain

$$\frac{df}{dt} - \frac{2}{3t}f = t^{-2/3}\,. \tag{15.9A.3}$$

We shall define $\boldsymbol{q}$ as the location (up to a numerical factor) of a point at the time when the free motion begins, i.e., at recombination. Hence

$$\boldsymbol{r} = t_r^{2/3}\boldsymbol{q}\,; \quad f(t_r) = 0\,. \tag{15.9A.4}$$

With this initial condition, we get the solution

$$f = 3t^{2/3}(t_r^{-1/3} - t^{-1/3})\,. \tag{15.9A.5}$$

This displacement of a particle in the comoving space is then given by the function ϕ satisfying

$$\phi = \frac{f}{t^{2/3}} = 3(t_r^{-1/3} - t^{-1/3})\,. \tag{15.9A.6}$$

As is apparent from the formula, $\phi \sim (t - t_r)$ close to the initial instant, whereas $\phi \approx 3t_r^{-1/3} = \text{const.}$ in the asymptotic limit $t \to \infty$. This differs substantially from the result $\phi \sim t^{2/3}$ as $t \to \infty$, a result which holds for growing perturbations that are intensified by their own gravitational field.

For the density, we find

$$\frac{\rho(q,t)}{\langle\rho(t)\rangle} \quad \mathfrak{Y}^{-1}, \quad \text{where} \quad \mathfrak{Y} = |d_{ik}|\,, \qquad d_{ik} = \delta_{ik} + \phi(t)\frac{\partial\psi_i}{\partial q_k}\,. \tag{15.9A.7}$$

In the vortex theory, the function $\psi(q)$ is by no means required to be of the potential form. Indeed the tensor $\partial\psi_i/\partial q_k$ is not symmetric, and its antisymmetric part corresponds to rotation. By a rotational choice of the coordinate axes for any given particle, however, one can bring the tensor $\partial\psi_i/\partial q_k$ locally, at a fixed q, to the form

$$\frac{\partial\psi_i}{\partial q_k} = \begin{vmatrix} \alpha & \omega & \delta \\ -\omega & \beta & \epsilon \\ \delta & \epsilon & -(\alpha+\beta) \end{vmatrix}. \tag{15.9A.8}$$

The quantity ω characterizes the rotation around the z-axis (the third axis). Because the initial motion is incompressible, the trace of the tensor vanishes:

$$\mathrm{Tr}\,(\partial\psi_i/\partial q_k) = \sum_i \partial\psi_i/\partial q_i = \alpha + \beta - (\alpha+\beta) = 0\,.$$

In addition to these freely specifiable diagonal entries though, the vortex-free (symmetric) part of the tensor is also characterized by the off-diagonal entries δ and ϵ. Having specified that the z-axis is the rotation axis, we are left with only one degree of freedom. This is the freedom in the rotation of the coordinates around the z-axis. We have made use of this to exclude the symmetric part of $\partial\psi_1/\partial q_2$ and $\partial\psi_2/\partial q_1$. The quantities α, β, δ, ϵ, and ω then depend on q and on the orientation of the axes.

In the notation adopted, we therefore obtain (exactly!)

$$\frac{\rho(q)}{\langle\rho\rangle} = \{1 - \phi^2[\alpha^2 + \alpha\beta + \beta^2 + \delta^2 + \epsilon^2 - \omega^2] - \phi^3[\alpha\beta(\alpha+\beta) + \delta^2\beta + \epsilon^2\alpha + \delta^2(\alpha+\beta)]\}^{-1}\,. \tag{15.9A.9}$$

As one might have expected, the expansion of $\delta\rho/\langle\rho\rangle = \rho/\langle\rho\rangle - 1$ for small ϕ begins with a term in $\phi^2 \sim (t \sim t_r)^2$. This is because the incompressible nature of the motion at the initial instant ($\partial\rho/\partial t = 0$ at $t = t_r$), a feature inherited from the prerecombination era, has been taken into account.

For certain values of components of the tensor, the corresponding motion can lead to the formation of a shock wave when the quantity in braces in the formula above goes to zero. In distinction to the case of potential-type motion (§ 14.2), this quantity never has three positive roots for ϕ. However, one positive root is quite possible even if $\omega \neq 0$. Even

then, though, the physical realization of every positive root for ϕ is not always necessarily a shock and an infinite density state. This follows if one keeps in mind the growth limitation that $\phi \to$ const. as $t \to \infty$.

In reality, however, the emerging density perturbations become gravitationally unstable when the growth of ϕ ceases (effectively $z \sim 0.3z_\tau$). These growing perturbations are of course the initial perturbations of the growing type of solution.

The nature of the expression for $\rho/\langle\rho\rangle$ (15.9A.9) forces one to assume that there is an anticorrelation between the density and the vorticity. In particular, at the point at which the components of the tensor satisfy $\alpha = \beta = \delta = \epsilon = 0$ and $\omega \neq 0$, the density decreases monotonically: $\rho/\langle\rho\rangle = (1 + \phi^2\omega^2)^{-1}$. At the point where $\omega = 0$ (no rotation), the density necessarily grows if $\phi \ll 1$. In asserting this, we are neglecting ϕ^3 in comparison with ϕ^2 and are recognizing that the coefficient of ϕ^2 in the brackets is positive. Thus, $\rho/\langle\rho\rangle \approx (1 - A\phi^2)^{-1}$ with $A > 0$ since $\alpha^2 + \alpha\beta + \beta^2 > 0$.

The statistical theory of the situation described here has not been worked out in detail.

16 RESEARCH ON PERTURBATIONS USING THE RELIC RADIATION

16.1 Introduction

A detailed study of the relic radiation (RR)—its spectrum and its angular distribution—can in principle give quite valuable information about deviations of the Universe from the ideal, homogeneous and isotropic cosmological model. There is no doubt that such deviations exist. Indeed, it was shown in the preceding chapter that deviations are necessary if the structure of the Universe—clusters of galaxies, galaxies and stars—is to emerge.

The determination of the deviations from homogeneity at a stage before the formation of celestial bodies, based upon the presently observed structure of the Universe, turns out to be difficult and fraught with ambiguities. Research on RR substantially supplements our knowledge of the perturbations. For example, small-scale perturbations decay even before hydrogen recombination occurs in the Universe and fail to leave a visible trace in the structure that subsequently emerges. In decaying, however, these perturbations liberate energy and distort the RR spectrum, thereby giving rise to deviations in the spectrum away from that described by the equilibrium Planck law.

On the other hand, the largest scale perturbations apparently have a small amplitude, so that statistical methods that concern themselves with discrete bodies are used in analyzing extragalactic objects (clusters, quasars, and radio sources). Yet this discreteness itself creates an additional source of "noise" or random errors. Therefore, for example, the precision of the determination of the Hubble parameter to this day is no better than 20%. The constancy of the Hubble constant determined with respect to objects in the northern hemisphere is also established with a precision of no better than 20%.

Meanwhile, measurements of the temperature of RR incident from various directions answer practically the same question: Is the expansion of the Universe the same in all directions, i.e., is the Universe isotropic? The precision of a measurement of the difference in the temperature in the most

The English translation of this chapter was edited by Mike Wilson, of the University of California at Berkeley.

convenient region, that of centimeter wavelengths, is better than 0.1%. At the present time (the beginning of 1972), observations have found neither spectral distortions nor any dependence of the temperature on direction. A positive result in the searches for deviations in the RR spectrum from that of equilibrium is lacking, notwithstanding the well-elaborated theory of such deviations with which observers are familiar.

At this point, an impatient reader can cease reading this chapter since it is concerned only with "nonexistent" effects. For two reasons, however, we shall continue the detailed discussion of the question. The first is that today's negative observational results still offer the possibility of giving limits on the possible amplitude of the perturbations. Indeed, such upper limits are rather interesting, especially in comparison of models with the observable structure. Moreover, a picture of the initial state of the Universe near the singularity, containing only quite small deviations from the idealized Friedmann model, can be sketched, given an upper limit. The second reason is that one can today expect further advances in the accuracy of the experimental methods. For example, simultaneous measurements in several spectral regions permit a more accurate separation of the RR fluctuations from those caused by nearby sources. One can thus hope that the effects considered below will actually be observed after an increase in the sensitivity by one or two orders of magnitude (to $\Delta T/T \sim 10^{-5}$). The theory set forth below will then become of practical importance in an interpretation of observations.

Recall that the spatially homogeneous situation was examined earlier. We explained at that time just what kinds of distortions arise as a result of the release of a given amount of energy at some particular period or, equivalently, at some value of the redshift z.

In § 16.2 and § 16.3 two specific reasons for the liberation of energy are examined—matter-antimatter annihilation (§ 16.2) and the transformation of the kinetic energy of peculiar motion into heat (§ 16.3). In § 16.2 it is shown just what limitations on the amount and distribution of antimatter follow from the absence of deviations in the spectrum. Analogous limitations on the peculiar motions are obtained in § 16.3. Actually in § 16.3, limitations on $\delta\rho/\rho$ are discussed. Some limits on peculiar velocity are discussed in § 16.5.

Subsequently, the question of the dependence of the RR temperature on direction in the presence of perturbations is studied. In § 16.4 a general scheme is given for calculating the temperature distribution for a specified peculiar motion of the matter and for a specified distribution of density pertubations. In a radiation field, which, to the first approximation, is of the equilibrium form and is homogeneous and isotropic, the problem is very much simplified. Under these circumstances, no deviations of the rays or focusing or defocusing of them can change the brightness. Only peculiar

motions, by means of the Doppler effect, can cause changes in the brightness temperature.

In the subsequent sections, this general scheme is applied to various types of perturbations. In § 16.5, it is applied to adiabatic perturbations, i.e., to potential-type motions with simultaneous changes in the density. These perturbations grow by virtue of gravitational instability, so their amplitude is now greater than ever.

Adiabatic perturbations also give rise to peculiar motion of that matter—that cluster or supercluster—of which our Galaxy, the Sun, and, in the final analysis, the planet carrying our observational apparatus, is part. RR permits one to determine at each point the system of reference relative to which the velocity of this peculiar motion (the random motion necessarily caused by perturbations) is measured. In § 16.5 is a discussion of the measurements of peculiar velocities and of the limitations following from the peculiar-velocity measurements on the amplitude of adiabatic perturbations.

Finally, we discuss in § 16.7 both vortex perturbations and long-wavelength gravitational waves as a possible source of the fluctuations in the RR temperature.[1] It is characteristic of these types of perturbation that they become weaker as expansion of the Universe proceeds. The effect therefore depends mainly on the situation at the earliest instant for observation—in this case the instant of recombination, when the plasma becomes transparent. In a definite interval of perturbation wavelengths, the effect depends strongly on the fact that the recombination of the plasma does not occur instantaneously. The radiation incident from a definite direction is in effect averaged with respect to the temperature of a thick emission layer.

Later, the secondary ionization of all of the primordial gas or of the part of it not entering into clusters leads to scattering and smooths out any possible angular dependence of the RR intensity. It would therefore seem that one can attribute the absence of observable perturbations to the smoothing action of scattering, even though the perturbations are initially large at the instant of recombination.

However, such a universal explanation—or rather excuse—is fraught with great difficulties. For an effective smoothing-out, a large optical depth is needed, which in turn requires an early secondary ionization (when $z \sim 50$–80). Yet the rate of heat transfer is so large at early times that it is difficult to keep the plasma hot and ionized. Also, the very processes that heat the gas introduce new perturbations into the relic radiation. In sum, the conclusion is that the observable high degree of RR isotropy is a strong argument in favor of the assertion that perturbations of the Friedmann homogeneous and isotropic model are small. In particular, this high degree of isotropy favors the theory of adiabatic perturbations and argues against the vortex (turbulence) theory of galaxy formation.

1. All questions related to gravitational waves and, in particular, to their possible role in causing RR fluctuations are covered in chapter 17.

16.2 Antimatter Annihilation

We shall now examine the annihilation of antimatter (antiprotons and helium antinuclei) and matter during the RD period. For the charge-symmetric cosmological model discussed in the literature, it is assumed that there are from the very beginning equal amounts of matter and antimatter in the Universe. Many authors consider such a model preferable on general philosophical grounds. The general principles behind such a preference will be discussed in Part V of the book, which is devoted to a study of the cosmological singularity. Here, we note merely that charge-symmetric or charge-asymmetric character for the model must be established at the singular instant itself, for this property remains quite unchangeable later.

In this section we shall find the requirements on the charge-symmetric model that are related to the distortion of the radiation spectrum caused by the inevitability of annihilation and the consequent liberation of energy.

A consideration of the RD period is important theoretically since one can compute the characteristics of the annihilation process at that time. The rate of the process is determined by the rate of diffusion of matter and antimatter to the boundary separating the respective regions.[2] In turn, the rate of diffusion depends on the friction arising from the motion of the electrons or positrons relative to the radiation. Furthermore, the protons cannot outstrip the electrons because the two are tied together by electrostatic forces. The theory of such a process is studied in detail in electrochemistry in connection with the diffusion of electrolytes. We shall here set forth the simplest variant of this theory, completely neglecting the friction of the protons moving through the radiation.[3]

The pressure of a completely ionized gas is evidently

$$P = (n_e + n_p)kT\,. \tag{16.2.1}$$

The gradient of this pressure is the force acting on the electrons in a unit of volume. The friction of the electrons in the radiation resists this force. Thus,

$$-\nabla P = F = n_e\alpha^{-1}v_e = n_e(4/3)\epsilon_r\sigma_T v_e/c\,, \tag{16.2.2}$$

where α is the electron mobility, σ_T is the Thomson cross-section, and ϵ_r is the energy density of the radiation ($\epsilon_r = 4 \times 10^{-13}z^4$ ergs cm^{-3}). The electrostatic interaction requires that $n_e = n_p$. From this equation, let us

2. We shall explain below the conditions necessary in order for diffusion to play the main role. To discuss these conditions, it is in particular necessary to find the value of the diffusion coefficient, which is therefore the first thing considered.

3. Allowing for a 7%–10% alpha-particle content in the primordial gas changes nothing. One can, in addition, verify that an earlier recombination of the helium also does not give a noticeable contribution because, in particular, the thermal velocity of helium atoms is 100 times smaller than that of the electrons. Curiously though, the annihilation of $\bar{p}$ and $\bar{n}$ could give quantities of D, T, and ^{3}He exceeding those produced by nucleosynthesis in the homogenous Universe.

find the average electron velocity v_e and the flux of electrons $q_e = n_e v_e$. We make use of the principle of electrical neutrality, $n_p = n_e$, to obtain that $v_e = -2\alpha kT\nabla(\ln n_e)$ and that the flux of protons identically equals the flux of electrons. Finally,

$$q_p = -2\alpha kT\nabla n_p = -\frac{3ckT}{2\epsilon_r\sigma_T}\nabla n_p = -\mathfrak{D}\nabla n_p . \qquad (16.2.3)$$

The coefficient $\mathfrak{D}$ in front of ∇n_p is by definition the diffusion coefficient. Substituting the expressions for ϵ_r and T that hold in the hot Universe, and also the numerical values of the constants, we get

$$\mathfrak{D} = 0.6 \times 10^{32} z^{-3} \text{ cm}^2 \text{ s}^{-1} . \qquad (16.2.4)$$

Now let us estimate in order of magnitude the distance l_z that a particle diffuses up to the instant when z achieves a given value. We use the formula $l_z = (6\mathfrak{D}t)^{1/2}$. Then, since $t = 3 \times 10^{19} z^{-2}$ s during the RD period, $l_z = 10^{26} z^{-5/2}$ cm. For a correct comparison, however, one must relate the path to a better unit—for example, the corresponding scale size today. Here one must of course also take into account the increase of this scale size by the factor $(z + 1) \approx z$.

The corresponding present-day scale size is thus $l = 10^{26} z^{-3/2}$ cm. With today's density of $\rho = 10^{-29}\,\Omega$ g cm^{-3}, the corresponding volume contains the mass

$$M = \rho R^3 = 10^{49}\Omega z^{-9/2} g = 5 \times 10^{15}\Omega z^{-9/2} M_\odot . \qquad (16.2.5)$$

Note that this formula is applicable only to the RD period before recombination. As an example, it gives $M = 3\,M_\odot$ when $\Omega = 0.1$ and $z = 1400$.

Let us now return to the general description of the phenomena. It is assumed in the calculation that the kinetics of the annihilation is determined entirely by diffusion. This means that the matter and antimatter nowhere coexist in noticeable amounts. All of space is thus divided into regions occupied by matter and regions occupied by antimatter. The annihilation takes place at the boundaries of these regions. Diffusion brings an amount of matter from one side to the boundary and an equal amount of antimatter from the other side to the boundary. This condition of equality of fluxes is achieved because the boundary between the two regions is not immobile. It moves toward that side from which the flux to an immobile boundary would be less than the oppositely moving flux.[4]

One can in this approximation compute the disappearance time for an "island" of antimatter surrounded by matter, such a disappearance being

4. The density of matter, n_B, and the negative of the density of antimatter, $-n_{\bar{B}}$, together form a continuous function of coordinates $f(r)$: $f(r) = n_B > 0$ in some regions, $f(r) = -n_{\bar{B}} < 0$ in other regions, while $f(r) = 0$ at the boundaries. The function $f = n_B - n_{\bar{B}}$ satisfies the equation $\partial f/\partial t = \mathfrak{D}\nabla^2 f$. From this it is easy to find the displaceplacement velocity of the boundary along the normal to the surface: $u \cdot \nabla f = -\mathfrak{D}\nabla^2 f$; $u_n = -|\nabla f|^{-1}\mathfrak{D}\nabla^2 f$.

due to the contraction of a boundary to a point. One can in addition compute the instantaneous fluxes of matter and antimatter moving toward the boundary, and hence the number of instances of annihilation per unit time per unit of area of the boundary. Such a picture, with a strict separation of particles and antiparticles, is merely a first approximation to reality. The annihilation itself bespeaks of the existence of a definite overlap region where particles and antiparticles are simultaneously present.

The situation just described is similar to that of the flame of a candle: Inside the flame there is an excess of hot fuel, while outside the flame there is an excess of oxygen. The flame itself is the boundary between the two environments (an oxidizing one and a reducing one) at which the chemical reaction proceeds. Now, by using the methods of the theory of combustion (Zel'dovich 1968), it is easy to give an estimate of the depth of the overlap layer and an estimate of the concentrations of particles and antiparticles in this region. One need only stipulate, as we shall do, that the distribution of the concentrations far from the overlap region be specified (from the first-approximation solution). Thus, let the equations $n = ax$ and $n' = 0$ for $x > 0$ and $n' = a|x|$ and $n = 0$ for $x < 0$ hold far from the boundary. Here, $x = 0$ is the middle of the overlap layer, and the coefficients on the left and the right are required to coincide. The annihilation rate is $q = \mathfrak{D}\partial n/\partial x = a\mathfrak{D}$ and is expressed in events $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The differential equations for diffusion then take the form

$$\frac{\partial n}{\partial t} = \mathfrak{D}\frac{\partial^2 n}{\partial x^2} - \beta n n' \; ; \qquad \frac{\partial n'}{\partial t} = \mathfrak{D}\frac{\partial^2 n'}{\partial x^2} - \beta n' n \; , \tag{16.2.6}$$

where β is a constant characterizing the annihilation rate and is expressed in $\mathrm{cm}^3\,\mathrm{s}^{-1}$. We seek a stationary ($\partial/\partial t = 0$) solution of these equations with the asymptotic form written above. From dimensional considerations, it follows that in the middle of the overlap layer at $x = 0$,

$$n = n' \sim (\mathfrak{D}a^2/\beta)^{1/3} \tag{16.2.7}$$

Meanwhile, the thickness of the overlap is

$$x_1 \sim (\mathfrak{D}/a\beta)^{1/3} \; . \tag{16.2.8}$$

It is then easy to convince oneself that the result

$$q \sim \beta n n' x_1 \equiv \mathfrak{D}a \tag{16.2.9}$$

also follows.

In order of magnitude, $\beta = \sigma v \sim 3 \times 10^{-16}\,\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$ for high-temperature annihilation, when no noticeable formation of $p\bar{p}$ "atoms," bound by Coulomb forces, takes place. This would increase the value of β if it were to occur. For an estimate, consider regions corresponding to a mass $M \sim 10^{12}\,M_\odot$ and to the critical density $\Omega = 1$. Correspondingly, $n(\infty) \approx 10^{-5} z^3\,\mathrm{cm}^{-3}$ in the region occupied by the matter, while n' has the same value in the region occupied by the antimatter. The characteristic size of the region is of

order $\lambda \sim (N/n)^{1/3}$, where N is the number of baryons in the region. We have $N \sim 10^{12} \times 2 \times 10^{43} \times 6 \times 10^{23} \sim 10^{69}$, so that $\lambda \sim 10^{25} z^{-1}$ cm. We take $a \sim n(\infty)/\lambda \sim 10^{-30} z^4$ cm^{-4} and $q \sim 10^{32} z^{-3}\, 10^{-30} z^4$ cm^{-2} s^{-1} $= 10^2 z$ cm^{-2} s^{-1}. Substituting these into the expressions presented above, we obtain $n(0) \sim 10^{-4} z^{5/3}$ cm^{-3} and $x_1 \approx 10^{26} z^{-7/3}$ cm. Finally, substituting $z \sim 10^4$, we find that $n(0)/n(\infty) \sim x_1/\lambda \sim 5 \times 10^{-5} \ll 1$. This also justifies the picture of the phenomenon used as a basis for the calculation.

The expression for the diffusion coefficient is valid in the limited region $10^8 > z > 1400$. For larger z, or higher temperatures, it is necessary to take into account two additional circumstances.

First, for temperatures higher than 1 MeV, a noticeable number of neutrons n is found in thermodynamic equilibrium with the protons p. Similarly, antineutrons $\bar{n}$ are found in equilibrium with antiprotons $\bar{p}$. Now n and $\bar{n}$ diffusion takes place more rapidly because the displacement of neutral particles does not give rise to a restoring electric field and therefore does not require e^- and e^+ displacement to compensate for this field. At temperatures lower than 20 MeV, when the equilibrium densities of the pions π^+, π^-, and π^0 are small, the diffusion coefficient for the neutrons is given by the formula $\mathfrak{D} = 5 \times 10^{33}\, \Omega^{-1} z^{-5/2}$ cm^2 s^{-1}.

Second, for temperatures higher than 3×10^8 K, there are e^+e^- pairs along with the electrons in thermodynamic equilibrium in the region occupied by the matter. The total concentrations of the protons, electrons, and positrons then satisfy $n_{e^+} + n_{e^-} \gg n_p$. In compensating for the electric field that arises due to the displacement of the protons, therefore, all of light, charged particles participate, so the diffusion coefficient for the protons is increased by a large factor.

With due regard for both circumstances, the diffusion and annihilation processes dispose of the small islands of antimatter by the instant when $z \sim 10^8$. These islands of antimatter are surrounded by matter (the case with opposite signs is also possible, of course), so these processes also serve to smooth out any gradients occurring over scales up to the order of 5×10^{11} cm, or, upon rescaling to present day dimensions, up to a scale of the order of 5×10^{19} cm. With respect to the present-day density, the corresponding characteristic mass is $\sim 10^{-3} \Omega\, M_\odot$. The liberation of energy when $z > 10^8$ is in no way expressed in the RR spectrum, because for these times the electrons and positrons accelerate the establishment of thermodynamic equilibrium.

What conclusion concerning a charge-symmetric Universe can one draw from these calculations? We shall take as a basis that clusters of galaxies and large galaxies, i.e., masses of the order of $10^{12}\, M_\odot$, consist of only one form of material—either matter or antimatter. Otherwise, active annihilation would now be taking place in the intragalactic gas and would give rise to a large number of energetic quanta ($E \sim$ 70–100 MeV) according to the reaction chain $p + \bar{p} \rightarrow \pi^0 \rightarrow 2\gamma$. But such quanta are not observed. There

nevertheless remains the possibility of a choice between two different forms of symmetric model that correspond to limiting cases. We shall not, however, discuss here all of the possible intermediate situations.

In the first case, regions with an excess of matter, B, and regions with an excess of antimatter, $\bar{B}$, both of the aforementioned size (corresponding to $10^{12}\,M_\odot$), are specified from the beginning—close to the singular state. The annihilation is minimal in this case. Why? Given a smooth density distribution, the annihilation rate is proportional to the square of the wave vector of the Fourier components of the distribution, i.e., it is proportional to $M^{-2/3}$. If a mass of the order of $3\,M_\odot$ is completely annihilated by the time of recombination, then a mass of the order of $10^{13}\,M_\odot$ is subject to partial annihilation at a rate that is $(3 \times 10^{12})^{2/3} = 2 \times 10^8$ times slower. What occurs after recombination? In the approximation of free motion for hydrogen atoms (or antihydrogen atoms $\bar{\mathrm{H}} = \bar{p}e^+$) with a thermal velocity, the characteristic path length is $l \sim u_0 t_r \sim 10^6\ \mathrm{cm\ s^{-1}} \times 10^{13}\ \mathrm{s} \sim 10^{19}\ \mathrm{cm}$. Moreover, the characteristic scale of a region containing $10^{12}\,M_\odot$ is of the order of $L \sim 10^{22}$ cm at the time t_r. Near the boundary, $n \sim n'x/L$ for a smooth distribution, so that the fraction of matter annihilated is of the order of $(l/L)^2$, i.e., 10^{-5} to 10^{-6}. For the later period, after secondary ionization, one can imagine that the gravitational and magnetic fields of astronomical objects that have already emerged hinder the annihilation.

In such a variant, the release of energy upon annihilation is not large; observations of the RR spectrum thus do not permit one to reject this variant. There are difficulties associated with it in other areas, however. By the time of recombination, the density fluctuations in this model are of order unity. Gravitational instability rapidly amplifies the fluctuations; the isolation of clusters occurs for a value of z not much smaller than the value at recombination. The density of clusters would then be of the order of the density of the matter at the moment of recombination, i.e., $\langle\rho\rangle(1400)^3 = 3 \times 10^{-20}\Omega\ \mathrm{g\ cm^{-3}}$. In reality, the density of the matter in clusters is smaller by several orders of magnitude. The variant is therefore untenable. It was Field who first raised this objection.

A second objection is more subjective, more a matter of taste. The interest in a charge-symmetric model is connected with the assumption of a close connection and correspondence between the laws of physics and the choice of the initial (singular) state for the cosmological model. The laws are charge-symmetric—hence symmetry for the model of the Universe as a whole is also preferable. But a natural question arises here: If the whole Universe is symmetric, then what destroys the symmetry of separate parts of it, in particular, of large parts for which $M \sim 10^{12}\,M_\odot$? Concerning this point, the second variant, which we shall now discuss, is more consistent.

Assume, therefore, that the B and $\bar{B}$ regions are small in the initial state, and that to make up for this the density of baryons (and antibaryons) is

correspondingly larger in these regions. A B or $\bar{B}$ excess in large regions for which $M \sim 10^{12}\, M_\odot$, is then a result of the fluctuations in the number and size of the small regions inside of each large region.

A possible reason for the emergence of small regions with an excess of matter or of antimatter is the supposed (according to the Omnès hypothesis; see § 6.3) separation of the hot matter of the neutral plasma into two phases. The calculations performed by Omnès and his collaborators, in which a large value is given to the motion of the phase boundary, lead to an isolation of regions with masses up to $10^{11}\, M_\odot$. In the eyes of the authors, however, this motion of the boundary is itself a consequence of the annihilation and energy release at the boundary. An estimate of the energy release averaged over a volume containing many regions gives $\Delta\epsilon \sim 20\epsilon_r$, where $\Delta\epsilon$ is the energy density released and ϵ_r is the radiation energy density, during the period $10^8 > z > 1400$ before recombination. This release, moreover, takes place mainly near $z = 1400$. But it is then absolutely intolerable, for, according to the theory of spectral distortions, a release $\Delta\epsilon \sim 0.1\epsilon_r$ already gives deviations from the Rayleigh-Jeans spectrum which exceed by a large factor the upper limit compatible with observations.

Finally, a more formal approach to the problem of antimatter is also possible in the realm of an asymmetric model. Assume that over the whole Universe there is on the average an excess of baryons corresponding to 10^{-8} of the number density of photons. Now we assume in the theory of entropy perturbations that $n_B(x, y, z) = n_{B_0}[1 + \delta(x, y, z)]$. By definition, $\langle\delta\rangle = 0$, so that there are in space regions with $\delta > 0$ as well as regions with $\delta < 0$.

For the formation of globular clusters, the root mean square value $\delta_{\text{rms}}(M) = [\langle\delta^2(M)\rangle]^{1/2}$ for the mass $M \sim 10^5\, M_\odot$ is important. Given the Fourier series, $\delta_{\text{rms}}(M) \sim [\delta_k{}^2 \cdot \frac{4}{3}\pi k^3]^{1/2}$, where k corresponds to the mass M and δ_k is the Fourier transform of $\delta(x, y, z)$. In order of magnitude, δ_{rms} $(10^5\, M_\odot)$ is between 0.01 and 0.1. Now the time of formation of the globular clusters depends on the value of this quantity. It is known further that for quite large masses, of the order of that corresponding to the size of the present-day horizon, the perturbations are small: $\delta_{\text{rms}}(10^{24}\, M_\odot) < 10^{-5}$. Thus, accepting the Dicke and Peebles hypothesis (1968) of the origin of globular clusters, we reach the conclusion that the spectrum $\delta(M)$ falls as M increases. The values of δ are given for the time of recombination; before this time, the entropy perturbations are frozen in.

Earlier, the values of $\delta_{\text{rms}}(M)$ for $M \ll 10^5\, M_\odot$ did not interest us since these perturbations do not grow after recombination. Now, however, we shall pose the question of whether δ_{rms} can be of order unity for small values of M. Given a regular variation of δ, this would signify that there are regions where $n_B = n_{B_0}\,(1 + \delta)$ becomes negative. It is necessary to regard such regions as regions occupied by antimatter. The frozen-in property existing during the RD period, i.e., the time independence, relates only to sufficiently large masses $M > M_\odot$.

If $\delta_{rms}(M) > 1$ for $M < M_\odot$, then such perturbations decay, and in addition the decay of the perturbations signifies the annihilation of the small regions with an excess of antimatter that are interspersed in the plasma, which has an overall excess of matter. Proceeding from the absence of distortion in the spectrum, Zel'dovich and Sunyaev (1970) obtained a definite bound on the value of $\delta_{rms}(M)$ in the interval $10^{-3} M_\odot < M < M_\odot$. Roughly $\delta_{rms} < 10$ at the lower edge of the interval and $\delta_{rms} < 0.5$ at the upper edge.[5]

If the existence of entropy perturbations with $M \sim 10^5 M_\odot$ is proved, then the limitations on δ_{rms} for small masses, and the absence of regions with antimatter will turn out to be especially important.

Given random occupation of the Universe by matter and antimatter during the early stages, B and $\bar{B}$ coexist in equilibrium for $T \geq m_p c^2$, so the appearance of regions with an excess of B seems inevitable. But another picture is also possible. Assume that the entropy is everywhere constant and that there is everywhere an excess of baryons originally. The motion of the matter then leads to the appearance of shock waves, to the growth of entropy, and to the appearance of entropy fluctuations. After the equilization of the pressure, the entropy s depends on the coordinates, the homogeneity is lost, and fluctuations are possible. But these fluctuations are of such a type that $n_B > 0$ everywhere, so that annihilation (for $T \ll m_p c^2$) does not occur. The observations, although insufficiently definite, persuade us to lean in favor of just such a picture of evolution.

16.3 Adiabatic Perturbations, Acoustic Oscillations, and Their Effect on the Relic Radiation Spectrum

It was explained in connection with the theory of galaxy formation that the perturbation amplitude before recombination required for galaxy formation is of the order of

$$\frac{\delta\rho}{\rho} \sim \frac{u\sqrt{3}}{c} \sim 10^{-4}. \tag{16.3.1}$$

This amplitude refers to perturbations whose wavelength corresponds to a mass of the order of $10^{13} M_\odot$. Perturbations with a smaller wavelength decay in the RD plasma as a consequence of photon viscosity and thermal conduction. The question examined below is whether one can observe these short-wave perturbations indirectly through distortions of the RR spectrum.

5. These bounds are related to the dependence on z of the amount of energy distorting the spectrum. By comparison with the result given in the original article, the lower limit of $10^{-3} M_\odot$ is changed when account is taken of the high-temperature diffusion (see above). At the upper boundary, there is appropriate regard made for the existence of regions for which $\delta(r) < -1$ even if $\delta_{rms} < 1$ in the Gaussian distribution.

As was explained earlier, one can observe an energy release[6] of the order of $3 \cdot 10^{-2}\%$ of ϵ_r for $1400 < z < 10^4$ and of the order of 100% of ϵ_r for $10^5 \lesssim z \lesssim 10^7$. Now the energy of oscillations equals twice their kinetic energy. In order of magnitude, the energy density is

$$\rho u^2 = \rho c^2 \tfrac{1}{3} (\delta\rho/\rho)^2 . \qquad (16.3.2)$$

Since the amplitude of acoustic waves near the decay threshold is $\delta\rho/\rho \sim 10^{-1}$, their energy is then $\epsilon \sim 10^{-8}\, \rho c^2 \sim 10^{-8} \epsilon_r$, i.e., a very small energy.

An observable effect is possible only in that case when the amplitude of the short-wave perturbations is quite large—of the order of $\delta\rho/\rho \sim 0.2$–0.3 for perturbations decaying when $1400 < z < 10^4$, and $\delta\rho/\rho \sim 1$ for perturbations decaying when $10^5 < z < 10^8$. In the linear theory of perturbations, decay during the former interval takes place for perturbations with masses satisfying $10^{12}\, M_\odot > M > 10^8\, M_\odot$, and in the latter interval for perturbations with masses satisfying $10^5\, M_\odot > M > 10^{-4}\, M_\odot$.

In reality, however, because the bounds on the amplitude are very weak, the amplitude may be significant. The linear theory of perturbation damping then turns out to be inapplicable.

What bounds can one *a priori* put on the perturbations without relying on the studies of the RR spectrum?

Density perturbations are connected with metric perturbations. But the order of magnitude of a metric perturbation in the period of growth, $t < t_r\, (M/10^{17}\, M_\odot)^{2/3}$, before the region of the duck's nose (see Fig. 43 and 44) equals that of density perturbations during the acoustic period, $t > t_r\, (M/10^{17}\, M_\odot)^{2/3}$, inside the duck's nose. But one must apparently exclude beforehand metric perturbations of order unity, because such perturbations can lead to the appearance of isolated, collapsing plasma structures. Indeed, for $\delta\rho/\rho \sim 1$, the total density of such "black holes," those of cosmological origin, would be gigantic. For an early collapse during the RD stage, moreover, it is necessary to take into account that photons would give the overwhelming contribution to the mass of such objects. We shall call these objects "otons" since they are predicted by GTR (and since the Russian language equivalent of GTR is OTO). The total density of black holes would be greater than the density of ordinary matter even when collapse and black hole formation only occur in a small part of the space (Zel'dovich and Novikov 1969). Hence GTR as applied to cosmology excludes black holes, so with certainty one can require that $\delta\rho/\rho < 1$.

Another bound on the amplitude of acoustic waves arises for a more prosaic reason, but is apparently more realistic. Acoustic oscillations were examined previously in terms of the linear theory, which is valid only for small perturbations. It was assumed in that analysis that the amplitude remains constant over the extended time from the moment of the beginning

6. We will not go into detail on this point here. Note, however, that all of the quantities calculated are brought to values appropriate to an average density parameter $\Omega \sim 0.3$.

of the oscillations (the lower boundary of the duck's nose) to the moment of dissipation (the Silk line). During this time, the small-scale waves, those with a small wavelength and a small period, manage to undergo many oscillations. Nonlinear effects, the transformation of sinusoidal waves into shock waves (Peebles 1970) and the generation of high harmonics, become substantial. Both effects are moreover continuously related to each other.

Now, a consistent and rigorous examination of the evolution of the "acoustic turbulence" for an aggregate of random acoustic waves is given in a study performed by Zakharov and Sagdeev (1969). It is, however, easy to solve the problem for a simplified model consisting of one-dimensional motion and without taking account of the cosmological expansion. Let us do this.

The propagation velocity for a specified state is

$$\frac{dx}{dt} = c + u = c_0\left(1 + \alpha\frac{\delta\rho}{\rho}\right) + \beta c_0\frac{\delta\rho}{\rho} = c_0\left(1 + \gamma\frac{\delta\rho}{\rho}\right), \quad (16.3.3)$$

where c_0 is the unperturbed sound speed and α, β, and γ are dimensionless coefficients of order unity. The sinusoidal wave in question transforms into a saw-tooth wave in the time when a distance equal to that between the crest and the trough of a wave, i.e., half of a wavelength, is traversed by a crest relative to a trough. This change occurs by virtue of the difference in the propagation velocities $c_0(1 + \gamma\delta)$ and $c_0(1 - \gamma\delta)$, where $\delta \equiv |\delta\rho/\rho|$. The time is of the order of $t_s = (\omega\delta)^{-1}$. During the next stage, the saw-tooth wave dissipates while propagating,[7] with an energy loss rate

$$-\frac{d\epsilon}{dt} \sim -\frac{d}{dt}(\delta^2\rho c_0^2) \sim \frac{\rho u^3}{\lambda} \sim \frac{\rho\delta^3 c_0^3}{\lambda} \sim \omega\rho c_0^2\delta^3 . \quad (16.3.4)$$

Here ρu^3 is the rate of energy dissipation per unit of surface area of the shock wave for a velocity change u in the shock wave. This equation, from which numerical factors have been systematically eliminated, is easily integrated. The result takes the form

$$\delta \approx \frac{\delta_0}{1 + \delta_0\omega(t - t_0)} < \frac{1}{\omega(t - t_0)} \quad (16.3.5)$$

Thus, if the wave undergoes a number of oscillations $N \sim \omega(t - t_0)$, the dimensionless amplitude must be less than N^{-1}, independent of the initial amplitude. Such are the results of the nonlinear theory. The time for the emergence of a shock wave from a smooth (for example, a sinusoidal) wave is of the order of $(\delta_0\omega)^{-1}$, which corresponds to the time for the amplitude to fall by a factor of 2. A more accurate account of the period required to

7. As is well known, the gradients in a shock wave automatically adjust themselves so as to yield the energy dissipation that follows from the conservation laws for coefficients of viscosity and thermal conductivity that are as small as is desired (see, e.g., Zel'dovich and Rayzer 1966).

establish the shock wave does not change this result that the total duration of the process is large.

What changes are introduced upon going to a more refined model? First, in accounting for the general cosmological expansion, one must keep in mind the changes in the wavelength and the frequency,[8] $\lambda \sim t^{1/2}$ and $\omega \sim t^{-1/2}$, and remember that the product $\omega(t - t_0)$ must be replaced by the integral $\int_{t_0}^{t} \omega dt$. For the lower limit, we have just $\lambda_0 = ct_0$ and $\omega_0 t_0 = 1$, so that

$$N = \int_{t_0}^{t} \omega dt = \left[\left(\frac{t}{t_0}\right)^{1/2} - 1\right] \approx \left(\frac{t}{t_0}\right)^{1/2} = \frac{z_0}{z}. \qquad (16.3.6)$$

Here $z_0 = 5 \times 10^8 \, (M/M_\odot)^{-1/3}$, and we choose z to be of the order of 10^5 to correspond to the instant when distortions of the spectrum first ensue. It is more difficult to take proper account of the propagation of the waves in all directions; the problem is indeed three-dimensional and statistical. Furthermore, given the nonlinear interaction of the waves, it is necessary to carry out the analysis in the language of harmonic generation.

An effective interaction arises only for waves acting for a long time and between which there is a constant phase relation. The projection of the velocity of one wave in the direction of another equals $c_0(1 + \gamma\delta) \cos \theta$, where θ is the angle between the wave vectors. We shall require that both corrections, $\gamma\delta$ and $(1 - \cos \theta)$, be of the same order so that a change in the speed of the wave is compensated by an inclination θ. From this we obtain the effective angles at which an interaction takes place:

$$\theta < \delta^{1/2}.$$

The concomitant solid angle thus satisfies

$$\frac{\pi\theta^2}{4\pi} \sim \delta. \qquad (16.3.7)$$

In another variant of the theory this ratio equals δ^2. We find correspondingly

$$\frac{d\delta^2}{dt} \sim -\omega\delta^{3+n} \qquad (16.3.8)$$

for the energy transfer to the higher harmonics, where $n = 1$ or $n = 2$. The solution of this equation is

$$\delta \sim (N + \delta_0^{-n-1})^{-1/(n+1)} < N^{-1/(n+1)}, \qquad (16.3.9)$$

and we have again omitted numerical factors.

Finally, we write the maximum energy which can go into spectral distortion for $z = 10^5$ and 10^6, given that the spectrum is left unchanged after

8. It is important here that the RD period is being considered. The expansion itself then does not change the amplitude of the wave.

the earlier period of nonlinear damping.[9] We shall moreover give the result for $n = 1$ and $n = 2$ and for $M = 10^4 M_\odot$ and $M = M_\odot$, though only for $z = 10^6$ in the latter case. With $n = 1$ and $M = 10^4 M_\odot$, we obtain $\epsilon' \delta^2 \sim 0.05$ when $z = 10^6$ and $\epsilon' \sim 0.005$ when $z = 10^5$; with $n = 1$ and $M = M_\odot$, we get $\epsilon' = \delta^2 \sim 0.002$ when $z = 10^6$. In the second situation, we have that with $n = 2$ and $M = 10^4 M_\odot$, $\epsilon' \sim 0.12$ when $z = 10^6$ while $\epsilon' \sim 0.03$ when $z = 10^5$; with $n = 2$ and $M = M_\odot$, we obtain $\epsilon' \sim 0.016$ when $z = 10^6$.

Thus, the conclusion is that with due regard for the nonlinear damping of the waves, distortions of the RR spectrum of the order of several percent could only be produced by waves corresponding to masses from 10^3–$10^4 M_\odot$ to $10^{12} M_\odot$ with a sufficiently large (larger than 0.2–0.3) initial amplitude. Smaller scale waves are weakened during an early period given any initial amplitude because of nonlinear effects. Moreover, perturbation amplitudes of order unity are excluded since they lead to the formation of black holes.

It has already been noted that one can apply the adiabatic theory of perturbations over the whole duration of evolution of the Universe from the singular state. It is then sufficient to specify metric fluctuations at the singular state. The density fluctuations can meanwhile be chosen equal to zero for $t \to 0$.

A flat spectrum for the metric fluctuations can be selected (Harrison 1970*a*; Zel'dovich 1973*a*) with an amplitude for metric perturbations of the order of 10^{-3} to 10^{-4} at all wavelengths. Then, the spectral perturbations correspond to an energy release of the order of $\Delta\epsilon/\epsilon = (10^{-6}$ to $10^{-8}) \ln (M_2/M_1)$, where M_1 and M_2 define the mass interval such that the corresponding perturbations decay during the "sensitive" period of the evolution, between $z_1 \sim 10^8$ and $z_2 = 1400$. We obtain $M_1 \sim 10^{-8} M_\odot$ and $M_2 \sim 10^{12} M_\odot$, respectively, so $\ln (M_2/M_1) \sim 4.5 \ln (z_2/z_1) \sim 50$. But even the large logarithmic factor cannot compensate for the small size of the original metric perturbations, whose amplitude appears quadratically. Thus, an energy release $\Delta\epsilon/\epsilon < 10^{-4}$ is obtained, which is of one or two orders less than the limit of sensitivity for detection of spectral perturbations.

All things considered, the adiabatic theory of galaxy formation agrees wholly with an absence of distortions in the RR spectrum.

16.4 Perturbations of the Spatial Homogeneity and Isotropy of the Relic Radiation

In accordance with the generally accepted picture, the Universe is to a first approximation homogeneous and isotropic—or in any case is so for a large part of its evolution. (Again, homogeneity means that local properties are independent of coordinates, while for any given point, isotropy means

9. We emphasize that for $\delta_0 > N^{-1/(n+1)}$ the energy remaining does not depend on the initial energy.

that the properties are independent of direction.) Indeed, the observed independence of the RR intensity, spectrum, or temperature with respect to direction confirms this picture most precisely. To this day, no deviations have been observed, even by precise experiments sensitive to deviations for which $\Delta T/T < 10^{-3}$.[10]

In the next approximation, though, one must take account of perturbations that destroy the homogeneity and isotropy. One must moreover presume that these perturbations can be of various types and range over all spatial scales from zero to infinity (or, in the case of a closed model, over all scale sizes from zero to the size of the Universe itself). Finally, the amplitude of the perturbations can vary with respect to the scale size.

From the viewpoint of cosmology—i.e., given the goal of learning about the initial state—perturbations with an amplitude for metric changes of $\sim 10^{-4}$ on a scale of 10^{12}–$10^{13}\,M_\odot$ are just as interesting as perturbations with an amplitude of $\sim 10^{-7}$ on a scale of $10^{24}\,M_\odot$ or as perturbations with an amplitude of 0.1 on a scale of $10^{-10}\,M_\odot$. One can then consider it "accidental" that perturbations with scale sizes from $\sim 10^{-4}\,M_\odot$ to $\sim 10^{12}\,M_\odot$ lead to $\delta > 1$ at the present time[11] and clearly manifest themselves in the observable structure of the Universe.

Small-scale perturbations, even if their initial amplitude is not small, decay long before recombination. They therefore do not manifest themselves directly as temperature fluctuations. Recall, though, that possibilities for indirect observations of small-scale fluctuations using the RR spectrum were discussed earlier in § 16.3.

Small-amplitude, large-scale perturbations manifest themselves at the present time[12] ($z = 0$) quite weakly as a small irregularity in the density and the velocity. We shall consider here the possibility of observing them with the help of RR. To do so, it is necessary to construct a theory describing the angular distribution of RR in a perturbed Universe.

Many researchers build such a theory by starting from the most general

10. In recent studies, authors claim an accuracy that is one or one and a half orders higher.

11. Not accidental is the emergence of life and of intelligence during the period when a definite part of the perturbations grow to amplitudes $\delta \gg 1$, i.e., to the nonlinear stage and to a point where celestial bodies form. Until then, $\delta < 1$ on all scales, so no bodies are in existence nor do the local, nonequilibrium conditions necessary for the emergence of nuclei heavier than hydrogen or helium yet exist; hence neither complex molecules nor life can exist until this time. The properties of the plasma at the time of recombination and the mass of $10^{13}\,M_\odot$ characteristic of the boundary of the region of decaying perturbations are also not accidental. Rather, they are determined by universal constants, by the properties of elementary particles, and by the value of the specific entropy.

12. If $\Omega < 1$, then small-amplitude perturbations do not grow, and never (even when $t \to \infty$) achieve large amplitude with respect to density, and hence do not lead to the formation of isolated, gravitationally bound bodies.

equation for the propagation of radiation in an arbitrary, time-dependent metric (see, e.g., Sachs and Wolfe 1967). This work is remarkable in that along with the complicated computations appears a clear and simple interpretation of the results, an interpretation we shall adopt below. In this book, nevertheless, the opposite procedure is implemented consistently: the simplest particular theories for the phenomena under consideration are constructed. At the price of a loss in generality, we tend toward simplicity and clarity, not only with regard to the interpretation but also with regard to the method of developing it.

What are the peculiar features of the problem that permit one to simplify the treatment of the phenomenon? In the absence of perturbations, RR would be completely isotropic; it would not depend on angle. In this situation, a change in the direction of rays—in particular, elastic scattering or focusing of them—does not give rise to observable effects. Indeed, this fact is summarized by a basic theorem of geometrical optics, which states that brightness (the flux of energy per unit of solid angle) is conserved given refraction or nonabsorptive scattering.

The theory turns out to be many times simpler than the theory that describes how large masses affect radiation from isolated sources. It is clear from what has been said that a gravitational lens at rest in no way at all destroys the exact RR isotropy for an Earthbound observer. More precisely, this applies to lenses at rest with respect to the general Hubble expansion and therefore not possessing any peculiar velocity.

On the other hand, a brightness change always occurs as a result of a redshift or blueshift of the spectrum. This is true whether we are concerned with the Doppler effect or with the change in frequency caused by a gravitational field. In both cases the frequency of all photons shifts by the same ratio, while the dimensionless photon occupation numbers for the cells of phase space do not change. As a result, a Planck spectrum remains a Planck spectrum, but with a temperature shift. In addition, this temperature shift can be different for different bundles of rays, depending on their "history"—on the velocity of the emitter, on that of the observer, on the gravitational potential at the points of emission and reception, and on the presence of nonstationary gravitational fields along the path.

The gravitational field of an isolated body does not destroy isotropy since the blueshift a photon undergoes along a path approaching the body is exactly compensated by the redshift it undergoes along the continuation of the path when it is receding from the body. Moreover, the gravitational potential of the cluster or supercluster to which our Galaxy belongs, together with the Sun and the Earth, does not play a crucial role in the propagation process. For this potential only gives rise to a general blueshift that is independent of direction and that an observer cannot decipher.

In a time-independent gravitational field, radiation remains in equilibrium at each point and, in particular, is isotropic, if it is in equilibrium at infinity;

only there is a general blueshift characterized by $\Delta\nu/\nu = -\phi/c^2$ in the Newtonian approximation. The laws of thermodynamics in GTR (see Landau and Lifshitz 1961) give the condition $(\sqrt{g_{00}})T = \text{const}$. Since $g_{00} = 1 + 2\phi/c^2$, this condition agrees with the Newtonian result presented above when due account is taken of the connection $T \sim \nu$ between frequency and temperature.

Thus, the possible RR anisotropy is completely characterized by the distribution $T(\theta, \phi)$ on the celestial sphere of the observed temperature characterizing the Planck radiation spectrum. Inasmuch as $T(\theta, \phi)$ completely determines the brightness,[13] there is no need to keep track of the twisting, contraction, or expansion of a bundle of rays received by an observer. Indeed, it is sufficient to keep track of the redshift along the path of a ray to determine $T(\theta, \phi)$. We emphasize that this shift depends on the Doppler shift as well as on the action of gravitational fields.

Photons are relativistic particles, so they are sensitive not only to the scalar potential (the difference between g_{00} and unity) but also to the other components of $g_{\alpha\beta}$. In particular, photons interact with gravitational waves, which do not exist in Newton's theory of gravitation.

However, one can always examine a ray in a locally inertial system at each point of its trajectory. Looked at in this way, the obvious effect of the metric (of the gravitational fields) on the redshift disappears. The calculation reduces thus to a summation (integration) of the Doppler frequency shifts along the entire trajectory. In short, the proposed method of calculation reduces to the following procedure. Imagine that space is filled with an aggregate of test particles on which no forces other than gravitational forces act. Examine the propagation of the rays in this medium and compute the frequency of the light as measured by observers moving together with the test particles. The frequency change over a short line segment depends only on the instantaneous relative velocity of neighboring particles and is calculated according to the Doppler formula. Over the line segment, the relative acceleration gives a contribution proportional to $(dr)^2$, but this contribution disappears upon integration since[14] $dr \to 0$.

13. In this section, the physical processes of absorption and emission of photons by a plasma and inelastic (e.g., Compton) scattering are not considered. As is shown in chap. 8, these effects can give rise to deviations from a Planck spectrum. Because, as observations show, all forms of RR perturbations are small, the principle of superposition of small perturbations holds. In this case, the principle signifies that one can study spatial (angular) perturbations of cosmological origin while considering the spectrum to be of the Planck form. To reiterate, this is in distinction to the situation when the perturbations arise from Compton scattering.

14. The relative acceleration $\Delta(d^2x/dt^2)$ between two neighboring particles is proportional to the distance Δx between them. An additional velocity difference, which depends on the acceleration, is proportional to the time for light travel between the particles. The contribution to the redshift that depends on the acceleration thus turns out to be proportional to $(\Delta x)^2$.

One should not get the impression that we are neglecting the influence of gravitational fields. They are taken account of accurately, although indirectly, through the velocity distribution of the test particles. In the unperturbed solution, the velocity field near any particle is given by an isotropic expansion law $\boldsymbol{u} = H(t)\boldsymbol{r}$. In the presence of longitudinal density perturbations or vortex or tensor perturbations (gravitational waves) the velocity field becomes anisotropic and is characterized by the law $u_L = H_{LK}r_K$, where H_{LK} is a tensor that depends on the time t as well as on the coordinate $\boldsymbol{r}$. Both the isotropy and homogeneity of the unperturbed model are simultaneously lost.

In essence this method reduces to determining the motion of the test particles accurately, if necessary, by using GTR. However, when the motion of the particles is found, the redshift can be computed in an elementary way. To use a proverb, we will not kill an ant with a sledgehammer twice.

There are additional considerations. For small-amplitude perturbations, both the change in the trajectory of a ray and the change in the redshift, ΔT, computed along the unperturbed trajectory are small. In the unperturbed field of particle motions, moreover, a change in the trajectory does not change the redshift. The contribution to the change in the redshift from the change in the trajectory thus involves the second order of a small quantity and is therefore not taken into account. These considerations are applied below to the calculation of $\Delta T(\theta, \phi)$, i.e., to the amplitude and angular distribution of the deviations from isotropy of RR.

We will be considering a Universe filled with matter satisfying $P = 0$, which is a good approximation after recombination[15] and the propagation of light rays therein. One must evidently consider only this postrecombination period since photons undergo repeated Thomson scattering from the free electrons before recombination. After recombination, however, the average photon energy—in particular, the energy of the photons in the Rayleigh-Jeans part of the spectrum—is many times less than the ionization energy of hydrogen, so the neutral gas is transparent to these photons. We shall discuss separately, in connection with specific assumptions concerning the perturbations, the possibility of secondary ionization of the gas and of photon scattering when $z < z_r$.

So, let us apply the principle set forth above to an arbitrary field of peculiar velocities in a dust medium for which $P = 0$; this same matter plays the role of the test particles. The basic equation is then

$$-\frac{1}{\nu}\frac{d\nu}{dt} = \frac{1}{c}\frac{du'_n}{dt}, \tag{16.4.1}$$

15. The radiation density equals 2×10^{-21} g cm^{-3} when $z = z_r = 1400$, whereas the matter density equals $3 \times 10^{-29}\Omega\,(H/75\ \text{km s}^{-1}\ \text{Mpc}^{-1})^2$ g cm^{-3}. Consequently, $P < (0.1)\rho c^2$ for $\Omega > 0.2$ if $H = 75$ km s^{-1} Mpc^{-1} and for $\Omega > 0.5$ if $H = 50$ km s^{-1} Mpc^{-1}. During the further course of the expansion, the approximation becomes still more accurate.

where u'_n is the projection of the velocity in the direction of a light ray, and dl is the differential path length in this direction. In the coordinate system with the primes, the particle at x_0 is at rest. The acceleration of this particle, relative to the light ray, is wholly determined from the spatial dependence of the velocity; thus

$$u'(x_0, t) = 0\,, \quad \frac{\partial u'}{\partial t}(x_0, t) = 0\,; \quad \frac{du'_n}{dt} = c\nabla_n u'_n\,. \tag{16.4.2}$$

Upon a transition to the "laboratory" frame, we add to $u'(x, t)$ the velocity of the particle at x_0, which depends only on time:

$$u(x, t) = u'(x, t) + u(x_0, t)\,. \tag{16.4.3}$$

Therefore,

$$\frac{du_n(x, t)}{dt} = \frac{du'_n}{dt} + \frac{\partial u_n(x_0, t)}{\partial t} = \frac{du'_n}{dt} - c\nabla_n\phi\,, \tag{16.4.4}$$

where ϕ is the Newtonian potential describing the motion (the acceleration) of the particles.

Upon integration, the first term gives simply the difference between the projections of the velocity at the beginning and at the end of the path:

$$\frac{\Delta T}{T} = \Delta \ln \nu = \frac{-u'_n(t_0) + u'_n(t_1)}{c}\,, \tag{16.4.5}$$

with t_0 and t_1 the times of emission and reception, respectively. This is the elementary contribution of the Doppler effect. Complications arise in the second term inasmuch as the potential depends not only on the coordinates but also on the time, and because the integrand is not a total differential.

If the perturbation wavelength is not large and the direction of the ray under consideration is not expressly perpendicular to the wave vector of the perturbation, then ϕ as well as $\nabla_n\phi$ are rapidly changing functions of time. If $\phi = \phi_0(t) \exp[i\boldsymbol{x}\cdot\boldsymbol{k}(t)]$ and if α is the angle between the ray and the wave vector, then

$$\nabla_n\phi = i\phi_0(t) \cos\alpha\, |\boldsymbol{k}(t)| \exp[i\boldsymbol{x}\cdot\boldsymbol{k}(t)]\,, \tag{16.4.6a}$$

so

$$\int \nabla_n\phi dt \approx \frac{1}{c}\int \phi_0(t) d[e^{i\boldsymbol{x}\cdot\boldsymbol{k}(t)}]\,. \tag{16.4.6b}$$

Here $\phi_0(t)$ and also $|\boldsymbol{k}(t)|$ are slowly changing functions of time—power-law functions, for example. The rapid variation of $\phi(\boldsymbol{x}, t)$ along the path of the ray results from the change in the exponent of $e^{i\boldsymbol{x}\cdot\boldsymbol{k}(t)}$ in the course of the propagation of the ray. The integral along the path of a ray, $\int_a^b \ldots dt$, is determined for a function of the above form principally by the contributions of the limits (see, for example, Zel'dovich and Myshkis 1972):

$$\int_a^b \phi(t)e^{i\omega t}dt = \phi\frac{e^{-i\omega t}}{i\omega}\Big|_a^b + \frac{(d\phi/dt)e^{i\omega t}}{\omega^2}\Big|_a^b + \ldots\,. \tag{16.4.7}$$

In our case, the slowly changing functions are proportional to powers of the time t reckoned from the singularity. Therefore, $d\psi/dt \sim \psi/t$, $d^2\psi/dt^2 \sim \psi/t^2$, and each successive term is smaller than the previous one by the factor $\omega t = kct \cos\alpha = ct \cos\alpha/\lambda$. For $\cos\alpha \sim 1$ and for short waves, there follows the possibility of retaining only the leading term. We then obtain

$$\frac{\Delta T}{T} = \frac{\Delta\nu}{\nu} = \frac{1}{c}[-u'_n(t_0) + u'_n(t_1)] + \frac{1}{c^2}[\phi(t_0) - \phi(t_1)]\,. \qquad (16.4.8)$$

In this approximation, the result has an extraordinarily simple and clear form: the change in the frequency is the sum of the Doppler shift corresponding to the peculiar velocity and the pure gravitational frequency shift.

The result is also presented in this form in the carefully executed work done by Sachs and Wolfe (1967) cited above as the culmination of rather complicated GTR calculations. However, in this study the gravitational shift contains the factor 1/3; it is still unclear how to interpret this coefficient classically.

The application of the principles we have set forth in the cases of large-scale adiabatic perturbations, vortex perturbations, and gravitational waves involves certain peculiar features each time. These features are discussed in the subsequent sections, along with concrete physical results.

16.5 The Detection of Density Perturbations Using the Relic Radiation

Consider density perturbations that are large in scale (from $10^{13}\,M_\odot$ to $10^{23}\,M_\odot$) and are small in amplitude. It is natural to assume that a sufficiently long time after recombination, we are concerned with growing perturbations. If, for example, growing ($\sim t^{2/3}$) and decaying ($\sim t^{-1}$) perturbations are produced at the time of recombination with the same amplitude, then by the present time the amplitude of the decaying perturbations turns out to be insignificantly small compared to the amplitude of the growing perturbations. From our concern with growing perturbations, two consequences then flow.

The first is a complete determination of the relation between the amplitude and phase of a density perturbation on the one hand and the amplitude, direction, and phase of the peculiar velocity on the other. For example, suppose

$$\begin{aligned}\frac{\rho - \langle\rho\rangle}{\langle\rho\rangle} &= \delta_{\boldsymbol{k}}(t)\,\sin\,[\boldsymbol{k}(t)\cdot\boldsymbol{x} + \psi_{\boldsymbol{k}}] \\ &= \left(\frac{t}{t_0}\right)^{2/3}\delta_{0\boldsymbol{k}}\sin\left[\left(\frac{t}{t_0}\right)^{-2/3}\boldsymbol{k}_0\cdot\boldsymbol{x} + \psi_{\boldsymbol{k}}\right],\end{aligned} \qquad (16.5.1)$$

where the coordinates are noncomoving and $\delta_{0\boldsymbol{k}}$, ψ, and $\boldsymbol{k}_0$ are constant quantities. The time t_0 is the present age of the Universe, 10^{10} yr. Finally,

we are considering a flat model with $\Omega = 1$ and $H = 75\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$. Then the peculiar velocity is

$$u = \tfrac{2}{3}\delta_{0k}\left(\frac{t}{t_0}\right)^{1/3}\frac{k_0}{t_0|k_0|^2}\cos\left[\left(\frac{t}{t_0}\right)^{-2/3}k_0\cdot x + \psi_k\right]. \qquad (16.5.2)$$

Thus, the present-day amplitude of the velocity, $|u|_{\max} = c\cdot|\delta_{0k}|\cdot[M/(4\times 10^{23}\,M_\odot)]^{1/3} = 4\times 10^2\delta_{0k}(M/M_\odot)^{1/3}\ \mathrm{cm\,s^{-1}}$ also corresponds to the present day amplitude of the density. The velocity is directed along the wave vector of the perturbation.[16]

The second consequence is that for a growing perturbation the maximum influence on RR occurs during the period nearest to the present, i.e., when the distance along the light path is of the order of a wavelength or half of a wavelength. For this reason it is also not essential whether we are concerned with adiabatic perturbations or with entropy perturbations before recombination. They would both produce only the growing perturbation mode in approximately equal amplitudes after recombination. One can also extend this principle to density perturbations appearing in the second (nonlinear) order from vortex perturbations, though here the first-order effect of the vortex perturbations on the density can turn out to be larger. Thus, we are studying matter motion that is on the RR background, a background which, to a first approximation, is homogeneous and isotropic. The motion is meanwhile specified by the formulae.

Now let us pass from consideration of a single wave to consideration of a statistical collection of waves. It is thereupon necessary to combine (vectorially!) the velocities due to perturbations of all wavelengths. Notice first the definition

$$\delta_{\mathrm{rms}}(M) = [\delta_k{}^2\cdot\tfrac{4}{3}\pi k^3]^{1/2}, \qquad (16.5.3)$$

where $M = \rho(\pi/k)^3$ is so constructed that

$$\langle\delta^2(x)\rangle = 4\pi\int_0^k \delta_k{}^2k^2dk = \int_M^\infty \delta_{\mathrm{rms}}{}^2(M)d\ln M\,. \qquad (16.5.4)$$

Here $\langle\delta^2(x)\rangle$ is the mean square deviation of the density from the average on a scale larger than a given lower limit M. From this we obtain the probable value of the velocity over large scales:[17]

$$\langle u^2\rangle^{1/2} = c\left\{\int \delta_{\mathrm{rms}}{}^2(M)[M/(4\times 10^{23}\,M_\odot)]^{2/3}d\ln M\right\}^{1/2}. \qquad (16.5.5)$$

16. At recombination, from which we are beginning the examination of the perturbation, the densities and velocities are not in the same relation that exists for a pure, growing perturbation. Originally there is a superposition of growing and decaying perturbations.

17. The limitation on the scale is essentially that the wavelength be less than ct or that the mass be less than $10^{23}\,M_\odot$. The large scales will be examined later; for them an increase in the RR alone is essential when the matter scale is also increased. In other words, a perturbation on a scale which exceeds the radius of the horizon, $\lambda > ct$, exists

Hence, the main perturbation of the RR isotropy observed by us depends on the present-day peculiar velocity of the observer himself! In this case, the resulting RR angular dependence is given by the very simple expression

$$\frac{\Delta T}{T} = \frac{\boldsymbol{u}\cdot\boldsymbol{e}}{c} = \frac{u}{c}\cos\alpha' , \tag{16.5.6}$$

where α' is the angle between the observation-direction $\boldsymbol{e}$ and the direction of the peculiar velocity $\boldsymbol{u}$.

Observations of the angular dependence of the RR intensity (its temperature) are ordinarily carried out with the help of an antenna fixed with respect to the Earth. With such a method, many possible experimental errors are eliminated; indeed, this is essential for the achievement of maximum sensitivity. Such an antenna draws a circle on the celestial sphere as the Earth rotates. Directly measured, therefore, is the dependence of the temperature on time. This dependence is decomposed into a Fourier series with a fundamental period of one day; the data from many days of observation are averaged together. The type of perturbation considered above, which depends on the present-day peculiar velocity, corresponds to a 24-hour harmonic:

$$\begin{aligned}\frac{\Delta T}{T} &= \frac{|u|}{c}\sin\phi\sin\psi\cos\left(\frac{2\pi t}{24\text{ hours}} + a\right)\\ &= A\cos\left(\frac{2\pi t}{24\text{ hours}} + a\right).\end{aligned} \tag{16.5.7}$$

Here ϕ is the angle between the velocity vector and the axis of rotation of the Earth, while ψ is the angle between the reception direction of the antenna and the rotation axis of the Earth.

The most accurate measurements give the result $|u|/c \approx 10^{-3}$. The troublesome feature of these measurements is that perturbations of diverse scale, combined with the motion of the solar system in the Galaxy and the motion of the Galaxy in the Local Group of galaxies, all give only a single resulting velocity vector. In the approximation under consideration, the angular dependence is such that one is not able to separate the contributions of the various enumerated causes from the single, total velocity vector.

Closely connected with this is a second difficulty. An accurate measurement of the RR isotropy has been carried out at only one point in the Universe. How much statistical weight should be ascribed to the result?

---

only in an open (and not in a closed) cosmological model. Even there, however, such a perturbation is not observable in the limit $\lambda \gg ct$ since this corresponds to a difference between the densities at the observation point and the region under consideration that is in principle unobservable. In order to get around this circumstance, one can either bound the upper limit of integration by $M_{23} \equiv M/(10^{23}\,M_\odot) = 1$, or replace M^2 in the integrand by $M^2\,[1 + (M/M_{23})^4]^{-1}$.

Might it not be that the value of $|u|$ (or of its upper limit $|u| < 300$ km s^{-1}) measured at the Earth is a result of an accidental compensation of the contributions of the various perturbations, namely those of the Galaxy and the Earth?

The answer to this question can only be given probabilistically. One can find the probability that the velocity of the Galaxy happens to be small, given a large average peculiar velocity. It is then quite essential to realize that the measurement of the velocity is in principle a measurement of three components of a vector and that these three components are statistically independent. Suppose then that the mean square value of this peculiar velocity obeys the probabilistic law

$$\begin{aligned} dW &= p(u_x, u_y, u_z)du_x du_y du_z \\ &= \left(\frac{2\pi u_0^2}{3}\right)^{-3/2} \exp\left[-\frac{3}{2}\left(\frac{u_x^2 + u_y^2 + u_z^2}{u_0^2}\right)\right] du_x du_y du_z . \end{aligned} \tag{16.5.8}$$

The probability that the velocity is less than u_1, where $u_1 \ll u_0$, is

$$W(u_1) \approx p(0, 0, 0)\frac{4\pi}{3}u_1^3 = \frac{4\pi}{3}\left(\frac{3}{2\pi}\right)^{3/2}\left(\frac{u_1}{u_0}\right)^3 \approx \left(\frac{u_1}{u_0}\right)^3 . \tag{16.5.9}$$

The large exponent of 3 is obtained precisely on account of the three-dimensional nature of the world—and the independence of the three components. With little loss of rigor, if a value $|u| < 10^{-3}c$ is measured on the Earth, one can say that the average value u_0 is less than $10^{-2}c$ with a probability of greater than 0.999, and the average value u_0 is less than $2.5 \times 10^{-3}c$ with a probability of greater than 0.93. In reality, we know the value of the projection of the velocity on the polar axis poorly today, since a large part of the measurements relate to the projection of the velocity on the equatorial plane.

In the limit of no information at all concerning the polar component, the estimates deteriorate. Nevertheless, we obtain $W(u_1) \sim (u_1/u_0)^2$, with the exponent 2 since there are only two components. In this case, $u_0 < 3 \times 10^{-2}c$ with a certainty of 0.999 while $u_0 < 4 \times 10^{-3}c$ with a certainty of 0.93, where these numbers apply under the same conditions as those given above. It is most advisable that the experimental difficulties be overcome so that the existence of RR isotropy or lack thereof can be established for the entire celestial sphere.[18] This observation is also important in connection with RR behavior in several anisotropic cosmological models.

The perturbation spectrum is therefore doubtless a falling one. Research on RR thus inflicts a decisive blow against the concept of a hierarchical structure, wherein it is assumed that galaxies form clusters, clusters form

18. A most important conclusion from the foregoing is that the present-day amplitude of density perturbations on the scale of the horizon is without doubt less than 0.01 in a flat, $\Omega = 1$, cosmological model. On the other hand, the present-day amplitude is much greater than unity on scales characteristic of clusters of galaxies.

superclusters, superclusters form supersuperclusters, and so on. This is because the amplitude of the density perturbations must be of the order of unity on each scale in this picture.

Let us make more precise the logic of this refutation. In addition to observations, the basis for RR isotropy also relies on general ideas concerning the evolution of a hot Universe from a dense period to the present day state. In the last analysis, though, these ideas too are based on observations and, foremost of all, on observations of the RR spectrum.

A direct study of the distribution of clusters of galaxies on the celestial sphere, carried out by Peebles and Yu (1970), also corroborates the absence of structural units corresponding to masses greater than $M \sim 10^{14} M_{\odot}$ and thus also contradicts the concept of hierarchical structure. However, RR research leads to the same conclusion with greater accuracy and reliability.

The relative velocities of distant objects (clusters of galaxies) are measured directly in terms of the shift of their spectral lines. However, a study of these measurements with the goal of obtaining the characteristics of perturbations is hampered by the knowledge that a large part of the measured effect is connected with the unperturbed Hubble redshift. In order to find the peculiar velocity, one must subtract the Hubble contribution, which is proportional to distance. It is evident that such a procedure is quite inexact and that the values obtained are rough. The error in the determination of the peculiar velocity corresponds to the inexactness with which the Hubble parameter is known ($[75 \pm 25]$ km s^{-1} Mpc^{-1} according to the data before 1972; $[53 \pm 5]$ km s^{-1} Mpc^{-1} according to Sandage's data [1972*a*, *b*, *c*]; see § 3.9).

Most difficult of all to contend with in the determination of H are the systematic errors. This is true in particular regarding the multistage system for defining the distance scale. Indeed, the dramatic leaps from $H = 560$ km s^{-1} Mpc^{-1} to the modern values have been related to just this scale measurement.

For the characteristics of the perturbations, it is sufficient to find the mean square deviation of the velocity from the main curve by confining the analysis to a uniform group of observations. The data from two such groups are given in Figure 52; they are from the article by Sandage (1972*a*, *b*, *c*). Ascribing all of the deviations from rectilinearity to peculiar velocities, we conclude that the root mean square deviation is about 3% of the Hubble velocity for the first group. For the second group, which consists of radio galaxies, the deviation is roughly twice as great. We get about $\langle u' \rangle \leq (0.04)$ $u_H = 0.04\, Hr$ over the whole interval $0.03 < z < 0.5$, i.e., for distances r from $0.03\, c/H$ to $0.5\, c/H$.

The density changes resulting from the growing perturbations are of the order of

$$\frac{\delta\rho}{\rho} \sim \frac{tu'}{\lambda} \sim \frac{u'}{Hr} \leq 0.04\,. \qquad (16.5.10)$$

Hence, we have obtained bounds on the size of the large-scale density fluctuations. For the largest scale, this bound is an order or an order and a half weaker than the bound which follows from a study of the RR anisotropy. However, we attach greater significance to the method involving the redshift of galaxies because many separate measurements (practically 100) are involved. On the other hand, the RR method of measuring the velocity involves only one object. It is therefore felt that the redshift method is less sensitive but more reliable, although this feeling is not easy to express numerically.

16.6 The Perturbation Spectrum and the Low-Density, Hyperbolic Model

RR measurements lead to a most important conclusion concerning the small size of the amplitudes of the largest wavelength perturbations: $\delta_{\text{rms}}(M) \leq 10^{-3}$ for $M \sim 10^{23}\, M_\odot$ and $\lambda \sim ct$; $\delta_{\text{rms}} \leq 10^{-2}$ for $M \sim 10^{20}\, M_\odot$ and $\lambda \sim 0.1ct$; and $\delta_{\text{rms}} \leq 10^{-1}$ for $M = 10^{17}\, M_\odot$ and $\lambda \sim 0.01ct$. Also, the relation $ct \sim 5000$ Mpc always holds for the above results. Finally, for $M \sim 10^{14}\, M_\odot$, we obtain $\delta_{\text{rms}}(M) \leq 1$. The density perturbation amplitude $\delta_{\text{rms}}(M)$ is here presented for the present time. But clusters of galaxies with masses of the order of 10^{13}–$10^{14}\, M_\odot$ doubtless exist and are rather boldly distinguished on maps of the sky today. Moreover, there are grounds for believing that these are formed not later than $z_{\text{clus}} \sim 2$ or $z_{\text{clus}} \sim 4$. This would signify that $\delta_{\text{rms}}(M_{\text{clus}}) \sim 1$ for this value of z_{clus} in the linear theory, and hence that the extrapolation to the present ($z = 0$) would give $\delta_{\text{rms}}(M_{\text{clus}}, z = 0) \geq 3$–$5$.

There is thus a contradiction between the estimates based on the 24-hour functions into which the RR distribution can be decomposed and those based on direct observations of the structure of the Universe.

The question of the reliability of the observations and of their statistical weight was discussed in the previous section. How reliable is the theory? Two basic assumptions were made, the first that the perturbations are small and the second that the unperturbed model is a flat one with $\Omega = 1$. If the perturbations are not small, then gravitationally bound bodies arise on the corresponding scales; we shall call them clusters.

In clusters, the velocities of the smaller constituents obey the virial theorem. The velocities thus do not grow in the course of time because the perturbations cease growing, though they do not decrease, as the expansion proceeds. After examining the formation of clusters, it is then necessary to consider large-scale perturbations within the confines of linear theory. With respect to them, the clusters can be thought of as atoms of a gas.

But still more important is the change in the theory flowing from the use of something other than the flat model. To understand why, it is only necessary to recall that the mass estimates of the matter in stars and in

gas invariably give values for the density lower than the critical value. Thus, a repudiation of the second assumption, that $\Omega = 1$, and an account instead of the features of the hyperbolic Universe for which $\Omega \ll 1$, apparently explain the relation between the velocity measured with the aid of RR (and according to the Hubble diagram) and the density perturbations. In a model with $\Omega \ll 1$ and in the linear theory, it is characteristic of perturbations that beginning at $z + 1 \sim 0.4\Omega^{-1}$, the peculiar velocity falls over the course of time, and density perturbations stop growing (see chap. 12).

Under this new assumption, therefore, one can harmonize the formation of clusters for which $\delta_{\text{rms}}(M_c) \sim 1$ when $z_c > 1$, with the small size of the velocity perturbations for $M > M_c$, where $\delta_{\text{rms}}(M) \sim 1$ when $z = z_c$. After the time when $z + 1 \sim 0.4\Omega^{-1}$, i.e., after the emergence of Milne expansion, the velocity falls inversely in proportion to the scale factor, and in proportion to $z + 1$.

An argument in favor of $\Omega \ll 1$ thus appears, an argument independent of the traditional methods for determining cluster masses, the gas density, or the curvature of the line on the Hubble diagram. This argument has much in common with Sandage's view. In a 1972 work, Sandage quite skeptically renounces the possibility of determining the deceleration parameter q_0, where $q_0 = \Omega/2$ for $P = 0$. He emphasizes (see also Tammann 1973*a*) that the deviations from the Hubble law would be quite large in regions with a density larger than the average if the average density $\langle \rho \rangle = \rho_c$ corresponds to the $\Omega = 1$ case. In short, what is observed is only weakly perturbed Hubble expansion!

Said differently, this argument coincides with the one set forth above, with the help of the properties of the perturbation-theory solutions.

16.7 The Angular Distribution of the Relic-Radiation Fluctuations

RR probably approaches Earthbound observers without undergoing any scattering after the $z = z_r \sim 1400$ plasma recombination. The angular distribution of the RR temperature therefore provides an opportunity to "see" directly (in the radio band) the prestellar plasma at an early stage of evolution. In particular, one can ascertain just how homogeneous the plasma is during this stage and just how precisely the velocity corresponds to that of a Friedmann model.

The local density and velocity deviations must produce local (small angle) temperature fluctuations. In principle, a measurement of the radiation from such fluctuations, which depends on the two angular variables θ and ϕ, yields much more information than does a measurement of the velocity of the observer with respect to the radiation field—a single quantity (more precisely, a single vector). The function $\Delta T(\theta, \phi)$ also characterizes

the amplitude and spatial scale of perturbations of the Friedmann model. The subsequent account of this is based on the work done by Zel'dovich and Sunyaev (1970); in the analysis here, however, we shall use the value 75 km s^{-1} Mpc^{-1} for H – not the value 100 km s^{-1} Mpc^{-1} which they use.

We begin with an explanation of the relations among the linear size of an inhomogeneity at the time of recombination, r_r, the size at the present time, r_0, the characteristic mass M, and the angle θ subtended by such an inhomogeneity. Taking, as everywhere, $H = 75$ km s^{-1} Mpc^{-1}, $\rho_c = 10^{-29}$ g cm^{-3}, and $P \ll \rho c^2$, and making use of the formulae of Part I of the book, we obtain the following relations:

In general, $\theta = r_r H_0/c\psi$, where $\psi(z \gg 1) = 2/\Omega z$; so

$$\theta = \frac{H_0 \Omega r_r z_r}{2c} = \frac{r_0 \Omega}{2.4 \times 10^{28}\ \text{cm}} = 1.3 \times 10^{-2} \Omega \left(\frac{r_0}{100\ \text{Mpc}}\right) = 45' \Omega \left(\frac{r_0}{100\ \text{Mpc}}\right). \tag{16.7.1}$$

On the other hand,

$$M = \Omega \rho_c \left(\frac{r_0}{2}\right)^3 = 1.5 \times 10^{16} \Omega \left(\frac{r_0}{100\ \text{Mpc}}\right)^3 M_\odot , \tag{16.7.2}$$

where $r_0 = z_r r_r = 1400 r_r$. Thus, we obtain,

$$\theta = 2.5 \times 10^{-3} \Omega^{2/3} \left(\frac{M}{10^{14}\ M_\odot}\right)^{1/3} \text{radians} \approx 10 \left(\frac{M}{10^{14}\ M_\odot}\right)^{1/3} \Omega^{2/3} \text{ minutes of arc} . \tag{16.7.3}$$

Silk (1968) first studied the question of whether the perturbations that lead to galaxy formation simultaneously give RR fluctuations. He gave an amplitude estimate corresponding to overall adiabatic contraction of the material and the radiation:

$$\frac{\Delta T}{T} = \frac{1}{3} \frac{\delta \rho_m}{\rho_m} = \frac{1}{3} \delta . \tag{16.7.4}$$

For an estimate of δ, perturbation growth is examined. If $\delta \sim 1$ when $z \sim 4$, then when $z = 1400$, $\delta \sim 1/300$. It is consequently expected (according to Silk) that

$$\frac{\Delta T}{T} \sim 10^{-3} \tag{16.7.5}$$

on a scale of several minutes of arc.

Searches for small-scale perturbations have occupied many radio astronomers. Note, for example, the recent works by Boynton and Partridge (1973) and Parijsky (1972). In the first, measurements were carried out at a wavelength of $\lambda = 0.35$ cm and it was shown that $|\Delta T/T| < 1.6 \times 10^{-3}$ on a scale of 1.3 minutes of arc. In the second investigation, the measurements were carried out at a wavelength of $\lambda = 2.8$ cm with the result that

$|\Delta T/T| < 3.0 \times 10^{-5}$ on a scale of 3 minutes of arc. Later, however (in a verbal communication), Parijsky changed his estimate somewhat. Upon accounting for the complex directional pattern of the antenna, he found that his data apparently do not contradict the result that $\Delta T \sim (1.3) \times 10^{-4}$ K on a scale of 3–5 minutes of arc. Even given this revision, however, the upper limit of the observable RR temperature fluctuations is at least three times lower than that estimated by Silk.

Does this signify that a perturbation-theory description of the formation of structure in the Universe is untrue? No, for in reality the Silk estimate requires serious alterations.

In accordance with the calculations done by Sachs and Wolfe (1967; see § 16.4), it is also necessary to take account of the temperature change during the course of the light propagation and the Doppler shift related to the plasma motion.

As long as $r_r < ct_r$ (which corresponds to $M \leq 10^{17} M_\odot$), a Newtonian analysis of the plasma is completely admissible at the time of recombination. Since the plasma can be described as a superposition of standing acoustic waves,

$$\langle \delta_r^2 \rangle = \langle (u_r/b)^2 \rangle \cdot \qquad \delta\phi_r = \frac{-4\pi G \rho_{m,r} \delta_r}{k_r^2} \cdot \tag{16.7.6}$$

$$\delta(z) = \frac{3 u_r k_r z_r t_r}{5(z + 2/5\Omega + \frac{3}{5})} \cdot \tag{16.7.7}$$

Here δ_r is the amplitude of the density perturbation, u_r is the peculiar velocity, $\delta\phi_r$ the gravitational potential perturbation, k_r the perturbation wave vector, t_r the age of the Universe, $\rho_{m,r}$ the density, and $b = c/(3 + 40\Omega)^{1/2}$ is the velocity of sound, all at the time of recombination. Finally, $\delta(z)$ is the amplitude of the density perturbation (calculated according to linear theory) at the time characterized by z. The formula for $\delta(z)$ is valid only when $z \ll 1400$; i.e., it relates to the period after recombination, when the perturbation is entirely in the growing mode. Transient processes occurring during the course of the recombination are accounted for in the expression for $\delta(z)$.

The relation between $\delta(z)$ and u_r describes roughly the results of linear perturbation theory for $\Omega < 1$. This relation permits one to determine the amplitude of the quantities of interest to us, δ_r, u_r, and $\delta\phi_r$, by specifying the time of formation of clusters or by specifying the present-day values for the inhomogeneity.

The fluctuations in the RR temperature consist of two parts:

$$\frac{\Delta T}{T} = \frac{u_r \cos\alpha}{c}; \qquad \frac{\Delta T}{T} = \frac{1}{3}\delta_r + \frac{\delta\phi_r}{c^2} \cdot \tag{16.7.8}$$

They are written separately because u and δ are not correlated in a standing wave: $u \sim \cos \boldsymbol{k}\cdot\boldsymbol{x}$, $\delta \sim \sin \boldsymbol{k}\cdot\boldsymbol{x}$, so $\langle u\delta \rangle = 0$. Meanwhile, δ_r and $\delta\phi_r$ are

correlated with each other in such a way that the gravitational redshift decreases the fluctuations that depend on the excess density.

Complete compensation is achieved when $k_r \sim (ct_r)^{-1}$, which holds at the edge of the mass interval under consideration. This edge coincides with the mass at the end of the duck's nose (see Figs. 43 and 44), i.e., with the Jeans mass $M_J \approx (1-3)\Omega^{-2}\, 10^{16}\, M_\odot$ for the plasma immediately before recombination, assuming $\Omega > 0.1$ (see § 11.1).

For $M \sim 10^{14}\, M_\odot$, one can neglect the term $\delta\phi_r/c^2$. Moreover, inasmuch as the sound speed is less than the speed of light, $u_r/c < \delta_r$ and the contribution to the temperature change of the Doppler effect is smaller than, although of the same order as, the "Silk" change in the temperature $\delta_r/3$.

For $M < 10^{17}\, M_\odot$, therefore, a detailed analysis confirms that the principal contribution is, in order of magnitude, $\Delta T/T \sim \delta_r/3$. But, there is a correction. In order to obtain the specified $\delta(z)$, the density perturbation δ_r must be less than that which Silk assumed, because the growing mode is excited by the motion of the neutral gas after recombination (see § 11.6). The correction is approximately $(M/M_J)^{1/3}$; i.e., it is of the order of 0.1 for $M = 10^{14}\, M_\odot$ and $\Omega = 0.3$. Thus, the expected value of $\Delta T/T$ is decreased to 10^{-4}.

The main effect causing the decrease in the expected value of $\Delta T/T$, however, is that recombination and the disappearance of the free electrons do not take place instantaneously. It was tacitly assumed above that the transition from an opaque plasma to an absolutely transparent neutral gas takes place instantaneously, so that one observes the velocity, temperature, and spectrum of a definite plasma particle. Given a gradual recombination, however, all of the quantities are averaged. The velocity, temperature, and density are, so to speak, weighted along the trajectory of a ray, with the weighting function given in such a way that it equals half of its maximum value within the limits $960 < z < 1135$ (see the previously cited references). During the time z changes inside these limits, a ray traverses a path length of $0.12ct$ given $\Omega = 1$. This is not many times less than the characteristic scale size $\chi = \lambda/2\pi = k^{-1}$ corresponding to the Jeans wavelength.

Fluctuations on a smaller scale are subject to very strong averaging and smoothing. According to the estimates of the authors, the coefficient describing the decrease in $\Delta T/T$ is 10^{-11} when $M = 10^{11}\, M_\odot$. The absence of small-scale fluctuations of the RR temperature therefore does not signify the absence of density, temperature, or velocity fluctuations in the post-recombination gas on the corresponding small scale. Note, though, that the smoothing effect completely disappears upon a transition to large-scale fluctuations of masses $M > 10^{17}\Omega^{-2}\, M_\odot$.

For a given density perturbation, the perturbation of the gravitational potential is proportional to the square of the wavelength, i.e., to $\sim M^{2/3}$. Hence, for $M > 10^{17}\Omega^{-2}\, M_\odot$, the gravitational redshift becomes dominant.

One can also formulate the result in a different way. For long waves we observe that

$$\frac{\Delta T}{T} \sim \frac{1}{3}\delta \approx \frac{1}{3}\frac{\delta\rho_m}{\rho_m}, \tag{16.7.9}$$

but it is necessary to take the value of δ at the time when the perturbation wavelength becomes equal to the radius of the horizon, $\lambda = ct$. For long-wave perturbations, this moment ensues long after recombination.[19]

Therefore, for the maximum mass, that corresponding to the entire observed Universe, the observed RR isotropy gives $\Delta T/T < 3 \times 10^{-4}$ and $\delta < 10^{-3}$. For perturbations whose characteristic size at the present time is of the order of $0.1ct = 500$ Mpc, we obtain the estimate $\delta < 0.1$ for the amplitude of their density today. This estimate is extraordinarily significant because one can then maintain reliably that no superclustering is expressed over 500 Mpc scales!

Until recently, the refinement of the observations has led to the observation of still larger structural units. The RR data place a clear limit on this process, however.

Finally, the study of fluctuations whose scale is much larger than that of the present-day horizon ct_0 remains insufficient. One can nevertheless attempt to explain the decrease in the observable RR fluctuations for perturbations on a superlong scale by the following reasoning. On such a scale, overall motion of the matter and of the radiation takes place primarily, while only to a small degree ($\sim ct/\lambda$) does motion of the matter relative to that of the radiation arise. It is of course only this relative motion which gives observable effects.

19. Recombination does not play a role for long-wave perturbations. The evolution of the material and of the radiation proceed, as before, in parallel and without separating until the time when $\lambda > ct$.

17 GRAVITATIONAL WAVES IN COSMOLOGY

17.1 Introduction

The gravitational waves predicted by GTR propagate at the speed of light. They are characterized by a definite energy density and momentum flux, or—for a chaotic superposition of waves—by a definite energy density and pressure. The distinctive feature of gravitational waves is their weak interaction with ordinary matter and with one another.

The recent sensational reports about the discovery of gravitational waves emitted by the nucleus of our Galaxy have apparently not been confirmed by later experiments. Nevertheless, these reports have stimulated great interest in various aspects of gravitational waves and, in particular, in the question of their cosmological role.

Does the Universe consist primarily of gravitational waves? Can the density of gravitational waves (the energy density divided by the square of the speed of light) at the present time exceed the density of ordinary matter or, at least, the density of the electromagnetic background radiation? Finally, what is the gravitational wave spectrum?

The weak interaction of such waves with matter does not permit one to give a direct experimental answer to these questions. In this chapter, therefore, we shall collect indirect information that bears on them. We shall assess the role of the gravitational radiation which might fill the Universe in terms of its various astrophysical effects. Let us first enumerate these effects:

1. The effect of the gravitational radiation on the overall expansion of the Universe and, in particular, on the age of the Universe.
2. The effect of the radiation (through the rate of expansion) on the perturbation growth law and the consequent connection with the theory describing the formation of clusters of galaxies.
3. The effect of the radiation on cosmological nucleosynthesis during the period when the temperature of the Universe is of order 10^9–10^{10} K. (The effect here is indirect, though the effect on the temperature is during an earlier stage of the expansion than pertains to points 1 and 2.)

The English translation of this chapter was edited by Dr. Bernard Carr, of Cambridge University and California Institute of Technology.

4. The effect of energy transfer from the waves to galaxies in clusters or to the clusters themselves.

5. The effect of the gravitational waves on the relic radiation. It is for long waves that this approach is most informative because the accuracy of the measurements is such that one can only show that the energy density of the gravitational radiation is cosmologically small in the long-wave region.

Theory also leads to several predictions concerning gravitational radiation. One ought to consider these points:

6. The possible existence of short-wave, thermal gravitational radiation analogous to the electromagnetic relic radiation. The important question here is whether thermodynamic equilibrium can be established near the singularity.

7. The connection between long-wave gravitational radiation and other types of metric perturbations—in particular, adiabatic density perturbations.

8. The generation of gravitational waves by stars and in the nuclei of galaxies.

Before carrying out such a complete examination of the questions connected with the cosmological role of gravitational radiation, however, one should recall the fundamental classical results concerning gravitational waves, namely:

9. The general expressions and numerical estimates of the standard[1] theory.

10. The results relating to the evolution of gravitational waves in an expanding Universe.

The 10 enumerated points essentially exhaust the contents of this chapter. The general conclusion, though not final, is rather pessimistic: a set of inequalities is established that limit the gravitational-radiation density. However, we stress that there are gaps in our knowledge connected with spectral regions which remain unstudied at the present time. In this chapter, therefore, we outline a program of further research and discuss methods for carrying it out. Note that the arrangement of the following sections does not correspond to the arrangement of the 10 aforementioned points.

17.2 General Information about Gravitational Waves

In empty space, for which $T_{ik} = 0$, a possible solution of the GTR equations is one with the metric form

$$ds^2 = c^2dt^2 - g_{\mu\nu}dx^\mu dx^\nu , \quad g_{\mu\nu} = \delta_{\mu\nu} + h_{\mu\nu}(t, x) , \qquad (17.2.1)$$

1. The account below expands upon the information on gravitational waves contained in Landau and Lifshitz's text (1973) and in Volume 1.

corresponding to a weakly perturbed Minkowski metric. Given this metric, the equation $G_{ik} \equiv R_{ik} - \frac{1}{2}g_{ik}R = 0$ leads, in the linear approximation, to equations for $h_{\mu\nu}$ of the form

$$\Box h_{\mu\nu} \equiv \left(\frac{1}{c^2}\frac{\partial^2}{\partial t^2} - \nabla^2\right) h_{\mu\nu} = 0 . \tag{17.2.2}$$

With the supplementary condition $h_{\mu\nu}^{(0)}k^\mu = h_{\mu\nu}^{(0)}k^\nu = 0$, the elementary solutions of these equations are

$$h_{\mu\nu} = h_{\mu\nu}(\boldsymbol{k}) = h_{\mu\nu}^{(0)}e^{i\boldsymbol{k}\cdot\boldsymbol{x}-i\omega t}, \qquad \omega = c|\boldsymbol{k}| ; \tag{17.2.3}$$

These elementary solutions form a complete system in the sense that any general solution can be expressed as a superposition of them, i.e., as the linear combination $\int h_{\mu\nu}(\boldsymbol{k})a(\boldsymbol{k})d^3k$. When using complex notation, it is, of course, understood that either $h_{\mu\nu}$ is the real part of the complex expression or $a(\boldsymbol{k}) = a^*(-\boldsymbol{k})$, a condition that guarantees the reality of the Fourier integral.

In particular, by superposing waves with parallel wave vectors $\boldsymbol{k}$, one can construct a solution of the form $h = h(\boldsymbol{n}\cdot\boldsymbol{x} - ct)$, where $\boldsymbol{n}$ is the unit vector $\boldsymbol{k}/|\boldsymbol{k}|$. In this solution the dependence of h on its argument is almost arbitrary—it is limited only by the transversality condition that $h_{\mu\nu}n^\nu = h_{\mu\nu}n^\mu = 0$ at every point in space.

Hence a gravitational wave propagates at the speed of light and gives rise to changes in the metric which are transverse with respect to the propagation direction. These changes will turn out to be important when we examine their interaction with matter. The metric (17.2.1) is synchronous because $h_{00} = 0$, $g_{00} = 1$, and $h_{0\alpha} = g_{0\alpha} = 0$. This permits one to readily determine the law of motion of one particular system of test particles in the field of such a wave. This solution is $x^\mu(t) = \text{const.}$, and it exhibits the feature that changes in both the mutual separations and relative motions of these particles are completely determined by the change in the metric coefficients.

By considering terms which are quadratic in the $h_{\mu\nu}$, one can calculate the stress-energy pseudotensor of the metric (Landau and Lifshitz 1973). The general expression for this takes the form

$$\begin{aligned} t_{00} &= \frac{1}{c}t_{01} = \frac{1}{c^2}t_{11} \\ &= \frac{c^2}{32\pi G}\left[\left(\frac{dh_{23}}{dt}\right)^2 + \frac{1}{4}\left(\frac{dh_{22}}{dt} - \frac{dh_{33}}{dt}\right)^2\right] . \end{aligned} \tag{17.2.4}$$

Here, the wave vector $\boldsymbol{k}$ is directed along the x^1-axis and, from the condition of transversality, only h_{22}, h_{23}, and h_{33} are nonzero. As Landau and Lifshitz emphasize, one can set the pseudotensor equal to zero at any point in space by means of a coordinate transformation. However, the integral of the stress-energy pseudotensor over a space which is flat at infinity has an

invariant significance as the energy-momentum $(E, \boldsymbol{p})$ four-vector of the gravitational field inside the volume under consideration. So long as the Euclidean condition is preserved at infinity, the four-vector $(E, \boldsymbol{p})$ does not depend on the choice of coordinates.

A packet of unidirectional, almost flat waves satisfies $|\boldsymbol{p}| = E/c$, i.e., its energy-momentum four-vector is almost null: $E^2 - \boldsymbol{p}^2c^2 = 0$. This corresponds to saying that the speed of gravitational waves equals the speed of light or that the rest mass of the graviton is zero. The qualification "almost" is related to the requirement that there may be wave-vector components, however small, in the directions perpendicular to the x^1-axis if the packet is limited in space in these directions.

If wave components with substantially different directions are added together, then the addition $\boldsymbol{p} = \boldsymbol{p}_1 + \boldsymbol{p}_2$ is vectorial while the addition $E = E_1 + E_2$ is performed as for scalars. The result is that $E > c|\boldsymbol{p}|$ or $E^2 - c^2\boldsymbol{p}^2 > 0$; the combination of waves moving in different directions thus has a nonzero rest mass.

In the case of a gravitational wave (one wave or an aggregate of waves) one can weaken the requirement on the form of the metric at infinity. It is sufficient to consider a volume, all scales of which are large in comparison with the gravitational wavelength. The volume average of the pseudotensor then possesses all of the properties of a true stress-energy tensor like the Maxwell stress-energy tensor for an electromagnetic field. This intuitively obvious property is introduced formally in Isaacson's (1968*a*, *b*) work by separating the metric into rapidly varying and slowly varying components.

For the elementary wave

$$h_{\mu\nu} = Re(\alpha_{\boldsymbol{k}} h_{\mu\nu}^{(1)} + \beta_{\boldsymbol{k}} h_{\mu\nu}^{(2)}) e^{i\boldsymbol{k}\cdot\boldsymbol{x} - i|\boldsymbol{k}|ct}, \tag{17.2.5}$$

where $h_{\mu\nu}^{(1)}$ and $h_{\mu\nu}^{(2)}$ are the two unit tensors transverse to $\boldsymbol{k}$ and $\boldsymbol{n} = \boldsymbol{k}/|\boldsymbol{k}|$, we obtain

$$\epsilon = \frac{c^2\omega^2}{32\pi G}(|\alpha_{\boldsymbol{k}}|^2 + |\beta_{\boldsymbol{k}}|^2) = \frac{1}{c} t_{0n} = \frac{1}{c^2} t_{nn}. \tag{17.2.6}$$

For a distribution of waves that is random with respect to direction (isotropic on the average or without a distinguished direction), the flux of energy goes to zero while the spatial components of the pseudotensor give an isotropic pressure:

$$\langle \epsilon_g \rangle = 3\langle P_g \rangle = \frac{c^2\omega^2}{32\pi G} \int (|\alpha_{\boldsymbol{k}}|^2 + |\beta_{\boldsymbol{k}}|^2) \frac{d\Omega}{4\pi}. \tag{17.2.7}$$

The quantities ϵ_g and P_g, the gravitational-wave energy density and pressure, along with the corresponding quantities for "ordinary" matter, enter into the right-hand side of the Einstein equations for the slowly varying part of the metric. We emphasize that ϵ_g enters into those equations with a factor G. It follows that, upon substituting ϵ_g and P_g, the gravitational

constant cancels. One would expect this because the coupling between the slowly varying and rapidly varying nonlinear metric components is a result of the nonlinearity of the equations $G_{ik} = 0$, and these equations contain no gravitation constant at all.

In the cosmological problem, for a model which is isotropic (on average) and filled only with gravitational radiation, we have the order-of-magnitude equation

$$-\frac{1}{a}\frac{d^2a}{dt^2} = \left(\frac{1}{a}\frac{da}{dt}\right)^2 = \omega^2\alpha^2 \approx \left(\frac{dh}{dt}\right)^2. \tag{17.2.8}$$

Here h characterizes any of the components $h_{\mu\nu}$, and α is the corresponding Fourier amplitude. Consequently, even a small amplitude

$$h \sim \frac{1}{\omega t} \sim \frac{H}{\omega}$$

for the metric perturbations can produce important cosmological effects. Here t is the age of the Universe and H is the Hubble parameter, $H = 10^{-18}\,\mathrm{s}^{-1}$. Thus for $\omega = 1\,\mathrm{s}^{-1}$, the value indicated for h is $\sim 10^{-18}$; for $\omega = 10^4\,\mathrm{s}^{-1}$ (the resonant frequency of Weber's detectors), $h \sim 10^{-22}$; and for $\omega = 10^{10}\,\mathrm{s}^{-1}$ (the average frequency of the relic radiation photons), $h \sim 10^{-28}$.

These small values for the dimensionless quantity h just characterize the weak interaction of gravitational waves with matter even when they play an important cosmological role. On the other hand, for the longest possible waves, those with a wavelength on the $\lambda \sim ct$, and a frequency $\omega \sim t^{-1} \sim H$, an amplitude of unity is necessary for a noticeable cosmological effect. Inasmuch as no dimensionless quantities differing from order unity are needed to characterize the effects of such waves, their effect on radiation incident from the horizon is also of the order of unity. We shall see below that isotropy of the relic radiation thus excludes the existence of gravitational waves with an amplitude of order of unity and a wavelength of order the horizon size.

Besides estimating h, which was done above, it is interesting to examine the quantities dh/dt and d^2h/dt^2. It is immediately apparent from equation (17.2.8) that dh/dt is of order the Hubble parameter if the gravitational waves dominate the density of the Universe and if the Universe is closed or flat ($\Omega \geq 1$). However, if $\Omega \ll 1$ and/or $\epsilon_g \ll \epsilon$, then $dh/dt \ll H$. The quantity d^2h/dt^2 is directly observable because the curvature tensor can be expressed linearly in terms of d^2h/dt^2.

A random (on average isotropic) collection of short-wavelength gravitational waves behaves in a Friedmann Universe as an ideal, ultrarelativistic gas with equation of state $P = \epsilon/3$. Consequently, the energy density of the waves changes as $\epsilon \sim [a(t)]^{-4} \sim (1 + z)^4$. For each separate mode the wavelength is proportional to the length scale of the Universe: $\lambda \sim a(t) \sim$

$(1 + z)^{-1}$; $\omega \sim (1 + z)$. Bearing in mind that $\epsilon \sim h^2\omega^2$, we thus have $h \sim a^{-1} \sim (1 + z)$; i.e., the wave amplitude, expressed in terms of dimensionless changes in the metric (with indices suppressed), decreases with time. Finally, note that the ratio of the wave energy in a given comoving volume to the wave frequency, an adiabatic invariant, remains constant during the evolution. This adiabatic invariant in some sense represents the number of quanta—in this case the number of gravitons—which are present. Thus, the number of gravitons is conserved during the adiabatic period of the expansion when the wave frequency is greater than the Hubble parameter and the wavelength is less than the horizon size. This is true for an individual wave as well as for a whole gas of waves. (For more detail about the evolution, see the next section.)

An isotropic random distribution remains isotropic in an isotropically expanding Universe.[2] However, given any global deviations from the homogeneity and isotropy of the Friedmann model, a most important feature of the graviton "gas" rapidly becomes apparent. This gas is *collisionless*, so the isotropy of the wave distribution is rapidly destroyed either by a global anisotropic expansion of the Universe or by any local anisotropic effects.

17.3 Gravitational Waves in the Theory of Small Perturbations of a Cosmological Model

In chapter 12 we examined Lifshitz's theory of small perturbations superposed on the homogeneous and isotropic Friedmann cosmological expansion. We noted there the possibility of invariantly classifying the perturbations by separating them into scalar (including density) perturbations, vector (vortex) perturbations, and tensor perturbations.

Tensor perturbations are defined in such a way that one cannot construct from them and the wave vector any nonzero scalar or vector quantities. Tensor perturbations therefore represent gravitational waves.

During the early stages of the expansion of the Universe, when the wavelength of the wave is greater than the horizon size, tensor perturbations represent "that which in the course of time transforms into a gravitational wave" or a "wave-to-be." Let us now consider in more detail the evolution of tensor perturbations and especially the transition of a wave-to-be into a "real" wave.

We shall first consider the relation between the horizon size and the gravitational wavelength. The horizon size[3] is of order ct. The wavelength is proportional to $a(t)$, so its instantaneous value is $\lambda = \lambda_0 a^{-1}(t_0)a(t)$, where $a(t_0)$

2. Given the additional condition that there is (on average) spatial homogeneity.

3. The radius of the Universe does not interest us because all features of the closed, open, and flat models are the same near the singularity. One can therefore consider a flat model with an infinite radius but with a finite horizon.

and λ_0 are the values at an arbitrary time t_0. As is well known, $a(t)$ grows more slowly than t: like $t^{2/3}$ for $P = 0$, $t^{1/2}$ for $P = \epsilon/3$, and $t^{1/3}$ for $P = \epsilon$. For each wave, therefore, one can find a moment of time t_1 such that, for $t < t_1$, $\lambda > ct$, while for $t > t_1$, $\lambda < ct$. For example, $t_1 = \lambda_0^2/c^2 t_0$ when $P = \epsilon/3$.

It is more convenient, however, to express quantities in terms of the comoving coordinates x, y, and z and the parameter η. In terms of these, the metric becomes

$$\begin{aligned} ds^2 &= dt^2 - a^2(t)[dx^2 + dy^2 + dz^2] \\ &= a^2(t)[d\eta^2 - dx^2 - dy^2 - dz^2]\,; \qquad d\eta = \frac{dt}{a(t)}\,. \end{aligned} \tag{17.3.1}$$

In this system of coordinates, the wave vector $\boldsymbol{k}$—the coefficient in the exponential $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$—is constant; so also is the comoving wavelength $\lambda_{\text{comov}} = 2\pi/|\boldsymbol{k}|$. The definition of the horizon is obvious in this system: since the equation for the propagation of light is $ds = 0$ or $d\eta = (dx^2 + dy^2 + dz^2)^{1/2}$, the distance to the horizon is simply η. Thus, the periods when $k\eta < 1$ and $k\eta > 1$ are naturally differentiated. The former is the initial period, when the distance to the horizon is less than the wavelength, while the latter is the period when the distance is greater. Since $k = \text{const.}$, whereas η grows with increasing t, there can be no doubt that $k\eta < 1$ *initially*, while $k\eta > 1$ *subsequently*.

For weak gravitational waves in a radiation-dominated Universe, Lifshitz obtained the solution

$$h_\alpha^\beta = \text{const.} \cdot G_\alpha^\beta \cdot e^{i\boldsymbol{k}\cdot\boldsymbol{x} + ik\eta}/a(\eta)\,, \tag{17.3.2}$$

where $a = \text{const.} \cdot \eta$. Moreover, the real and imaginary parts are solutions separately. From such solutions and their complex conjugates, one can construct two linearly independent solutions:

$$\begin{aligned} h_\alpha^{(1)\beta} &= C_1 G_\alpha^\beta \cos(\boldsymbol{k}\cdot\boldsymbol{x} + \phi_1)\,\frac{\sin k\eta}{k\eta}\,; \\ h_\alpha^{(2)\beta} &= C_2 G_\alpha^\beta \cos(\boldsymbol{k}\cdot\boldsymbol{x} + \phi_2)\,\frac{\cos k\eta}{k\eta}\,. \end{aligned} \tag{17.3.3}$$

Of these, the first remains finite near the singularity at $\eta = 0$. Consequently, finite or small metric perturbations at the singular state lead to finite or small-amplitude gravitational waves at late times when $k\eta > 1$.

For any particular wave, with given values $C_1 < 1$ and $C_2 = 0$, the gravitational-wave energy during a late stage[4] is a fraction of order $\alpha = C_1^2$ of the energy of the matter. (If the matter consists of an ultrarelativistic gas of photons, the ratio $\alpha = \epsilon_g/\epsilon_r$ remains constant in the course of the ex-

4. But still radiation dominated.

pansion.) We emphasize, however, that it is in principle inadmissible to ask about the density of gravitational waves at such an early stage that $k\eta < 1$.

We have been discussing the evolution of an *individual* wave. For small amplitude metric perturbations and correspondingly small-amplitude waves, the principle of superposition holds. Therefore, the above results generalize naturally to a random superposition of waves. The condition that the perturbations be finite near the singularity imposes a definite limitation on the randomness. Note that this condition is already accounted for in equation (17.3.3). Indeed, it turns out that this condition distinguishes standing gravitational waves. Take the most general linear combination of complex solutions with given values $|\boldsymbol{k}|$ and with given directions $\pm \boldsymbol{k}/|\boldsymbol{k}|$:

$$f = \frac{C_3}{k\eta} e^{i\boldsymbol{k}\cdot\boldsymbol{x} - ik\eta} + \frac{C_4}{k\eta} e^{-i\boldsymbol{k}\cdot\boldsymbol{x} - ik\eta} + \frac{C_5}{k\eta} e^{i\boldsymbol{k}\cdot\boldsymbol{x} + ik\eta} + \frac{C_6}{k\eta} e^{-i\boldsymbol{k}\cdot\boldsymbol{x} + ik\eta} . \tag{17.3.4}$$

The condition that this combination be real requires that $C_4 = C_5{}^*$ and $C_6 = C_3{}^*$. Writing $C_3 = Ae^{i\phi}$, $C_6 = Ae^{-i\phi}$, $C_4 = Be^{i\psi}$, and $C_5 = Be^{-i\psi}$ with A, B, ϕ, and ψ real, we then obtain

$$f = \frac{A}{k\eta} \cos(\boldsymbol{k}\cdot\boldsymbol{x} - k\eta + \phi) + \frac{B}{k\eta} \cos(-\boldsymbol{k}\cdot\boldsymbol{x} - k\eta + \psi) . \tag{17.3.5}$$

This expression represents two traveling waves moving in opposite directions and with arbitrary and unequal amplitudes. Such a solution is real and satisfies the wave equation, though near the singularity its amplitude is (almost everywhere) infinite. Now, in order to formulate the limitation on the amplitude of f as $\eta \to 0$, we transform equation (17.3.5) to the form

$$f = \frac{\cos k\eta}{k\eta} [A \cos(\boldsymbol{k}\cdot\boldsymbol{x} + \phi) + B \cos(\boldsymbol{k}\cdot\boldsymbol{x} - \psi)] + \frac{\sin k\eta}{k\eta} [A \sin(\boldsymbol{k}\cdot\boldsymbol{x} + \phi) - B \sin(\boldsymbol{k}\cdot\boldsymbol{x} - \psi)] . \tag{17.3.6}$$

The limitation on f requires that $B = -A$ and $\psi = -\phi$, so we are left with

$$f = 2A \frac{\sin k\eta}{k\eta} \sin(\boldsymbol{k}\cdot\boldsymbol{x} + \phi) . \tag{17.3.7}$$

The amplitudes for the two oppositely moving waves are equal, so that together they form a standing wave. Recall that the same conclusion was obtained for acoustic waves (scalar waves accompanied by a density change) and led to Sakharov's prediction for the amplitude of oscillations after recombination. For gravitational waves, however, no method has yet been devised, even in principle, for differentiating between traveling and standing waves. Note that the nodal surfaces defined by $\boldsymbol{k}\cdot\boldsymbol{x} + \phi = 0, \pm 2\pi, \pm 4\pi, \ldots$ are different for the various wave vectors in a random superposition.

17.4 The Expected Intensity of Relic, Short-Wavelength Gravitational Radiation

We shall now consider several variants of the scenario in which a background of relic gravitational radiation is expected. A general characteristic of these variants is the choice of the time $t = t_{\rm Pl} = 10^{-43}$ s as the earliest moment when our analysis applies.

The simplest variant results from the assumption that at the time $t_{\rm Pl}$, thermodynamic equilibrium holds between the gravitons and the other elementary particles and antiparticles. It is also assumed that the number of kinds of particle is finite and that the interactions among them are weak. At a temperature higher than the rest energy mc^2 of all of the particles, we have an ultrarelativistic gas, and this is the situation we shall consider. We shall do so by analogy with the way Peebles considered the question of relic neutrinos in the Universe.

Suppose that the energy density of all of the kinds of particles, including gravitons, is $\epsilon = \mathfrak{N}\sigma T^4$ for $T > mc^2$, where σ is the constant appropriate for electromagnetic radiation and where $\mathfrak{N}$ is approximately equal to the number of kinds of particles. The corresponding graviton entropy is $4\sigma T^3/3$, while the entropy for the remaining particles is $(\mathfrak{N} - 1)4\sigma T^3/3$. In the course of the expansion, the graviton entropy per unit comoving volume is conserved. But, with the exception of a small number of protons, all of the heavy particles and antiparticles eventually annihilate: in addition to the gravitons (which are being accounted for separately) at a time when $T \sim 5$–10 MeV there remain only electrons, the two kinds of neutrinos, their antiparticles, and photons. These leptons inherit the entropy of the heavy particles, so T_l, the lepton temperature, and T_g, the graviton temperature, are related by

$$(1 + \tfrac{7}{4} + 2 \times \tfrac{7}{8}) \times \tfrac{4}{3}\sigma T_l^3 = (\mathfrak{N} - 1)\,\tfrac{4}{3}\,\sigma T_g^3\,. \qquad (17.4.1)$$

In the subsequent evolution, the photons and the electron-positron pairs decouple from the neutrinos. When the e^+e^- pairs annihilate, their entropy is inherited by the photons and the radiation temperature is thereafter related to T_l by

$$T_r^3 = \tfrac{11}{4}\,T_l^3\,. \qquad (17.4.2)$$

Comparing temperatures, we thus find that

$$T_g = \left(\frac{4}{11} \times \frac{9}{2} \times \frac{1}{\mathfrak{N} - 1}\right)^{1/3} T_r\,. \qquad (17.4.3)$$

Putting $\mathfrak{N} \approx 20$–40 and $T_r = 2.7$ K, we obtain $T_g = 1$ K for the present-day graviton temperature. Note that, in this variant, we can always neglect the contribution of the gravitons to the total density since their energy density is only of order 0.01 times the radiation density.

Of course, we are interested not only in the total gravitational-radiation density but also in its spectrum or frequency distribution. Indeed, the effects of gravitational waves on other forms of matter and the methods for detecting gravitational waves are completely different for different frequencies. This should be borne in mind as we proceed with the exposition.

In the preceding discussion, where we examined gravitational waves initially in thermodynamic equilibrium with all other forms of matter and radiation, we concluded that the gravitational-wave wavelength in the spectral region where the maximum energy is concentrated is of order a fraction of a centimeter. The same assertion holds for the relic electromagnetic radiation. We now continue the examination of short-wave relic radiation on the assumption that equilibrium conditions do *not* exist at the moment t_{Pl}.

Two extreme assumptions are: (1) that there are no gravitational waves at all at the initial time and that the matter is hot and has normal entropy; (2) that the energy of the short-wavelength gravitational waves specifies the overall density almost entirely. The resultant state today evidently depends on the relaxation rate or of subsequent physical processes, or, more precisely, on the relation between the relaxation and expansion rates. A calculation shows that at the Planck time the product of the relaxation rate,[5] $\sigma n u$, and the characteristic expansion time, t_{Pl}, is of order unity. Indeed the transformation of two gravitons into a pair corresponds to the Feynman diagram of Figure 62, consequently, the cross section is proportional to G^2. In a system of units where $\hbar = c = 1$, the dimensions of G are cm^2. Hence, in order to obtain a cross section, it is necessary to multiply by the square of the energy, viz.,

$$\sigma = G^2 E_g^{\,2} . \tag{17.4.4}$$

Here we are using the fact that the rest masses of all of the particles, including the newly created ultrarelativistic ones, cannot enter the result. In the more familiar form, with the requisite factors of $\hbar$ and c, we have

$$\sigma = G^2 E^2 c^{-8} , \tag{17.4.5}$$

but henceforth we use units in which $\hbar = c = 1$.

At the moment $t_{\mathrm{Pl}} = G^{1/2}$, the corresponding energy density is $\epsilon = G^{-2}$. The equilibrium spectrum of gravitational waves (assuming they contain all of the energy-assumption [2] above) then corresponds to a temperature such that

$$\epsilon_g = T^4 = G^{-2} , \qquad T = G^{-1/2} . \tag{17.4.6}$$

This means that the energy of individual gravitons equals $G^{-1/2}$, which is the energy corresponding to the Planck mass; while their density is $n \equiv$

5. We think of relaxation as a particle collision involving the creation or annihilation of gravitons. The collision cross section is σ, the particle number density is n, and the particle speed is u.

$G^{3/2}$. The corresponding cross section is $\sigma \sim G^2/c^2 \sim G$. Hence, with $n = c^{-3} \sim 1$, we find that the total number of pairs created at the time t_{Pl} is $n' \sim \sigma n^2 t_{\mathrm{Pl}} = G^{3/2} \sim n$. Note that this calculation has been carried out crudely—without dimensionless factors and with the replacement

$$n'(t_{\mathrm{Pl}}) = \frac{1}{V(t_{\mathrm{Pl}})} \int_{t_{\mathrm{Pl}}}^{t_1} \sigma(t) n^2(t)\, V(t) dt \to \sigma(t_{\mathrm{Pl}}) n^2(t_{\mathrm{Pl}}) t_{\mathrm{Pl}} . \quad (17.4.7)$$

This replacement can only be wrong by a dimensionless factor, and it is necessitated by the fact that the integral diverges like a power law at the lower limit if one replaces t_{Pl} by 0. The quantity n' is determined by the value of the function at t_{Pl}. The conclusion is that the number of created pairs is of order the initial number of gravitons.

If we examine the scenario based on assumption (1), according to which gravitons are absent at the instant t_{Pl}, it is evident that the inverse process will produce a number of gravitons comparable to the number of other particles. One can hence expect that the energy density of short-wave gravitational waves will turn out in both cases to be of order the equilibrium value, i.e., in accord with the temperature of the remaining matter and around 1 K today. The difference in the energy densities predicted by the first and second assumptions amounts to a factor of perhaps 5. The difference does not, however, involve factors of the type $Gm^2/\hbar c$ which might in principle have entered into the problem.

It was noted above that, if there is equilibrium at t_{Pl}, then as a consequence of the annihilation of particles with nonzero rest mass, the present-day density of gravitational-wave energy should be $\epsilon_g \approx 0.01\epsilon_r$. One can suppose that the result $\epsilon_g < \epsilon_r$ at the present time remains valid even if $\epsilon_g \gg \epsilon_r$ at t_{Pl} because the situation would then change through relaxation to one of equilibrium. On the other hand, again on account of relaxation, it is doubtful that one could have $\epsilon_g < 10^{-3}\epsilon_r$ today even if $\epsilon_g \equiv 0$ in the initial state. (In this connection, see the stipulation below.)

The presence of gravitons, even in small amounts, can turn out to be important if the expansion is anisotropic. This is because gravitons become collisionless earlier than neutrinos. (For more about this, see Part IV.) Finally, recall the limiting assumptions under which the foregoing results apply: (1) Only short-wave gravitational waves have been considered (long waves will be examined separately later); and (2) we have not considered those cosmological models in which matter is initially cold and the entropy accumulates at a time t_e long after t_{Pl}. In this case (see Part V), the ratio ϵ_g/ϵ_r turns out to be small:

$$\frac{\epsilon_g}{\epsilon_r} \sim \left(\frac{t_e}{t_{\mathrm{Pl}}}\right)^{-r} \sim \left(\frac{Gm^2}{\hbar c}\right)^{+q} \quad (17.4.8)$$

with $r > 0$ and $q > 0$.

17.5 The Equipartition Theorem and Long-Wavelength Gravitational Radiation

In the theory of small perturbations of the homogeneous, isotropic cosmological model, gravitational waves are formally independent of other types of perturbations; moreover, gravitational waves with different wave vectors are independent of one another. Here, independence signifies that the differential equation for the amplitude of a given wave does not contain the amplitudes of other waves or other types of perturbations. In other words, the variables separate and the initial amplitude of each wave can be given arbitrarily—independent of the other amplitudes. Nevertheless, one can draw plausible conclusions concerning the expected amplitudes provided one specifies a definite picture for the initial state.

The homogeneous and isotropic Friedmann solution does not precisely describe our Universe. This is apparent from the existence of structure manifested by galaxies, for example. On the other hand, the theory of nucleosynthesis in homogeneous, isotropically expanding matter accords well with observations. It is therefore natural to require that the true solution possess the property that during an early stage it behaves locally as a Friedmann solution.

If one assumes that the $g_{\mu\nu}$ are close to their Friedmann values near the singularity, but that the differences of the various components from those values are independent of one another, then one draws a conclusion known as the "equipartition hypothesis." This hypothesis says that metric perturbations related to density (scalar) perturbations are equal in amplitude to metric perturbations related to gravitational (tensor) waves (Zel'dovich and Novikov 1972).

We shall now consider the spectral characteristics of the quasi-isotropic solutions; i.e., we shall separate the perturbations into plane waves. Spatial isotropy immediately implies that any perturbation amplitude depends only on the magnitude and not on the direction of the wave vector. The equipartition hypothesis speaks about the equality of two functions of $|\boldsymbol{k}|$—the amplitudes of the scalar and tensor perturbations. Consider the time when the wavelength becomes equal to the horizon size, i.e., when $\lambda = ct$ or $k\eta = 1$ (see § 17.3). According to the hypothesis, at this time the relative density perturbation amplitude has the same order of magnitude as the metric perturbation, i.e.,

$$\left(\frac{\delta\rho}{\rho}\right)_k = h_k\,, \qquad k\eta = 1\,. \tag{17.5.1}$$

Hence, one can express the amplitude of the gravitational wave in terms of the amplitude of the density perturbation.

For a scale corresponding to clusters of galaxies (a mass $M \sim 10^{13}\,M_\odot$ or

a present linear scale of $\lambda \sim 30$ Mpc), we infer a gravitational wave amplitude of $\sim 10^{-4}$. Thus

$$\frac{dh}{dt} \sim 10^{-4} t^{-1}, \qquad \rho_g \sim 10^{-8}\langle\rho\rangle = 10^{-8}\rho_r . \tag{17.5.2}$$

Here we have passed from separate Fourier amplitudes to quantities integrated over a single section of the logarithmic wavelength scale. The quantity ρ_g is ϵ_g/c^2, where ϵ_g is the energy density of the gravitational waves; and ρ_r is the radiation density. For the scale considered, the time when $k\eta = 1$ occurs during the radiation-dominated era, when ρ_r is practically the same as the total density. The ratio ρ_g/ρ_r itself remains constant right up to the present time. We thus estimate the present energy density of gravitational waves whose scale is that of clusters of galaxies as

$$\begin{gathered}\rho_g(M \sim 10^{13}\, M_\odot) \sim 10^{-8}\rho_r = 10^{-42}\ \mathrm{g\ cm^{-3}}\ ; \\ \epsilon_g \sim 10^{-21}\ \mathrm{ergs\ cm^{-3}}\ ; \quad \frac{dh}{dt} = 10^{-24}\ \mathrm{s^{-1}} = 10^{-6}H\ .\end{gathered} \tag{17.5.3}$$

On the scale over which density perturbations are known, therefore, the gravitational waves are only weak.

Over still larger scales, the amplitude of gravitational waves is limited by the isotropy of the electromagnetic relic radiation (see § 17.7). Over scales smaller than those characterized by a mass of $10^{13}\, M_\odot$, however, it is difficult to observe such perturbations.

Fortunately, this is a scale on which the equipartition hypothesis is useful. Density perturbations decay before recombination, and the heat released by their decay must change the spectrum of the relic radiation. But such changes are not observed and this allows one to impose the limit $\epsilon_{\mathrm{acous}} < 0.05\epsilon_r$ for the mass interval $10^{13}\, M_\odot > M > 10^5\, M_\odot$. The equipartition hypothesis then implies $\epsilon_g < 0.05\epsilon_r$ for the same mass interval.

According to the equipartition hypothesis, the existence of a flat spectrum of scalar metric perturbations, such as can explain both the structure and the entropy of the Universe (see § 24.9), would imply that gravitational waves only make a small contribution to the total energy density over all wavelengths from millimeters to 10^{10} pc. However, the existence of such a flat spectrum has not been demonstrated.

17.6 The Generation of Gravitational Waves in the Present Epoch

The generation of gravitational waves goes on in galactic nuclei, quasars, and binary star systems and through the activity of pulsars. Although connected with cosmology only indirectly, these phenomena are worthy of mention if only because it is desirable to collect information concerning the origin of gravitational waves in one place.

Gravitational waves of noncosmological origin were examined in Volume 1. Much attention focused in particular on Weber's experiments to detect them. Estimates were made of the energy of a gravitational-wave packet necessary to excite Weber's detector, and we noted the difficulties involved in explaining Weber's results given what is known about the relevant astrophysics. Indeed the situation seemed inexplicable at the time Volume I was written, although there were no other experimental data.

It now seems firmly established that the "peaks" and "coincidences" seen by Weber are not caused by gravitational waves. Using similar antennae but changing the method of recording, Braginsky and his co-workers have *not* found statistically significant coincidence. Tyson (1973) has come to the same conclusion. In short, both experiments indicate an absence of Weber-type outbursts of gravitational waves.

To estimate the amount of noncosmological gravitational radiation, therefore, it is necessary to turn to the theory of wave generation and to free oneself from the allure of the first, unconfirmed experimental data.

We shall begin with general energy considerations. In all cases except the collision of two black holes, the release of energy in the form of electromagnetic radiation is greater than that in the form of gravitational radiation. This assertion applies to the collisions of ordinary stars, white dwarfs, and neutron stars, and to processes involving pulsars. Hence the energy density of gravitational waves with such an origin is less than the energy density of noncosmological (nonrelic) electromagnetic radiation, i.e., it is substantially less than 10^{-13} ergs cm^{-3}. This corresponds to $\rho_g < 10^{-34}$ g cm^{-3}.

For the collision of two black holes, on the other hand, the gravitational radiation reaches several percent of the rest mass if there is relative orbital motion. For a head-on collision, it is less than a percent. Gravitational radiation can also play a significant role in the evolution of clusters containing a large number of black holes (see Zel'dovich and Podurets 1964, 1965). But such a cluster becomes unstable when the gravitational mass defect becomes several percent.[6]

Only a part of this energy loss goes into gravitational radiation, however. Consequently, one can immediately draw an important conclusion: noncosmological gravitational radiation cannot close the Universe! For if the total mass of ordinary matter and of black holes of all sizes is insufficient to close the Universe, then the gravitational radiation emitted by such bodies can only change the total density slightly. Moreover, if black holes constitute, say, 10% of the matter density, and if in turn 10% of the mass of the black holes is converted into gravitational radiation, then the radiation density can only be 1% of the matter density, i.e., $\sim 10^{-32}$ g cm^{-3}.

Let us now recall the arguments one can use to place limits on the density

6. In the most advantageous case, that of a disk, the mass defect reaches 40 percent (Bardeen and Wagoner 1971).

of the weakly interacting particles. The same arguments are also useful in the context of gravitational radiation because all the limits considered below depend on the fact that any form of matter and energy influences the overall dynamics of expansion of the Universe.

The age of the Universe, the density of the matter in it, and the Hubble parameter are related to one another. Chapter 2 presents the formulae for calculating the density. Let us turn to these formulae. Although the age of the Universe (from the moment of the singularity) is not known exactly, we can place a lower limit on it. With high probability, the age is greater than $(10\text{-}12) \times 10^9$ yr and, with absolute confidence, it is greater than 5×10^9 yr. The last figure is the age of the solar system. From this one obtains an inequality for the density: $\rho < 3 \times 10^{-29}$ g cm^{-3} if $H = 75$ km s^{-1} Mpc^{-1}, and $\rho < 5 \times 10^{-29}$ g cm^{-3} if $H = 50$ km s^{-1} Mpc^{-1}. These values apply if $P = \epsilon/3$. Note that they exceed the density of the electromagnetic radiation by a large factor and in this sense they are not very interesting for our present purposes.

Another limit is related to the indirect effect of weakly interacting particles, gravitational waves in particular, on the development of density perturbations. For details one should refer to the work done by Guyot and Zel'dovich (1970), but let us recall the essence of the argument: density perturbations grow in a stationary Universe according to the law $\delta\rho/\rho \sim e^{\omega t}$, where $\omega = (4\pi G\rho_m)^{1/2}$. Generalizing this law to the case of an expanding Universe in the natural way suggests the alternative expression $\exp(\int \omega dt)$. Agreement with the exact calculation requires multiplying the exponent by a factor of $\sqrt{(2/3)}$. The expansion law depends on the density of the gravitational waves. We have

$$\rho_m = \rho_{m,0}(1+z)^3 = \rho_c\Omega_m(1+z)^3 ;$$
$$dt = \frac{1}{H_0}\frac{dz}{(1+z)^2[1+\Omega_m z + (\Omega_r+\Omega_g)z(2+z)]^{1/2}} , \tag{17.6.1}$$

where Ω_m is the ratio of the density of nonrelativistic matter (that contained in galaxies) to the critical density $3H_0^2/(8\pi G)$, and Ω_r and Ω_g are the same quantities for the electromagnetic radiation and gravitational waves. After substituting the expressions for ω and ρ_c and then integrating from recombination to the present (the period over which the perturbations grow), we obtain

$$\alpha \equiv \sqrt{\left(\frac{2}{3}\right)} \int \omega dt$$
$$= \int_{z=0}^{z=z_r} \frac{\Omega_m^{1/2} dz}{\{(1+z)[1+\Omega_m z + (\Omega_r+\Omega_g)z(2+z)]\}^{1/2}} . \tag{17.6.2}$$

For $\Omega_m = 1$ and $\Omega_r = \Omega_g = 0$, this gives the well-known result

$$\alpha = \ln(z_r + 1) , \quad e^\alpha = z_r + 1 = 1401 . \tag{17.6.3}$$

Consider now the effect of the additional terms. It is well known that $\Omega_r = 10^{-4}$, and we shall set $\Omega_m = 0.2$ as an example. If $\Omega_g = 0$, the term involving Ω_r has almost no effect and

$$\alpha \equiv \alpha_0 = 5.6\,, \qquad e^{\alpha_0} = 270\,. \tag{17.6.4}$$

If we set $\Omega_{g,1} = 100\Omega_r$ and $\Omega_{g,2} = 1000\Omega_r$, however, we obtain a significant decrease in the perturbation growth:

$$\alpha \equiv \alpha_1 = 2\,, \quad e^{\alpha_1} = 10\,; \qquad \alpha \equiv \alpha_2 = 1\,, \quad e^{\alpha_2} = 3\,. \tag{17.6.5}$$

We know that clusters of galaxies form during the present epoch or perhaps somewhat in the past, when $z \sim 4$. For this to occur, the requisite initial perturbations are larger for smaller values of e^{α}. In the case $\Omega_{g,1} = 10^2\Omega_r$, therefore, the density perturbations at recombination must reach the value $\delta\rho/\rho \sim 0.1$. For adiabatic perturbations this corresponds to $\delta T/T \sim 0.03$. It is probable, however, that such fluctuations in the relic radiation are excluded by present observational data.

In general, the theory behind this question is insufficiently worked out to make reliable estimates. The principal problem is that in the course of the discussion one is obliged to make many assumptions, and this detracts from the conclusiveness of the argument.

Let us turn finally to an estimate of the gravitational-radiation density that is based upon nucleosynthesis considerations. Recall the chain of argument given by Shvartsman (1969). The amount of ^{4}He forming through cosmological nucleosynthesis depends upon the neutron to proton ratio at the freeze-out time. By taking account of the temperature dependence of the rate for the reactions

$$n + e^+ \rightleftarrows p + \bar{\nu}\,, \qquad n + \nu \rightleftarrows p + e^-\,,$$

Shvartsman obtained the relations

$$T_f = T_{0f}\left(1 + \frac{\Omega_g}{4.5\Omega_r}\right)^{1/6}, \qquad \frac{n_n}{n_p} = \exp\left(-\Delta m c^2/kT_f\right). \tag{17.6.6}$$

Here T_f is the freeze-out temperature, T_{f_0} is the value of T_f if $\Omega_g = 0$, and Δm is the neutron-proton mass difference. If $\Omega_g = 0$, they imply $T_f = T_{0f} = 10^{10}$ K and $n_n/n_p = 0.15/0.85$, so $\sim$30% by mass of the matter becomes ^{4}He. Setting $\Omega_g/\Omega_r = 10$ and $\Omega_g/\Omega_r = 50$, we obtain $(n_n/n_p) = 0.24$ and $(n_n/n_p) = 0.32$. The respective amounts of ^{4}He are then 40% and 50%. It is likely that a more detailed examination, particularly for small values of Ω_m such as $\sim$0.1–0.2, would also give rise to an increase in the deuterium and ^{3}He concentrations.

Now it was assumed in the above calculation that the number of neutrinos and antineutrinos are equal. If the symmetry $n_{\nu_e} = n_{\bar{\nu}_e}$ did not apply, then the number of variants of the picture would be enormous. We are thus considering only three methods for estimating the gravitational radiation

density. We emphasize that it is the average gravitational field of the gravitational radiation that is important in these analyses. The effect is therefore quadratic in the amplitude h of the wave.

Can one choose from these three limits the most sensitive one, i.e., the one giving the best upper bound on Ω_g? No, for such an approach would be incorrect. Not only does each method lack complete reliability and make use of additional assumptions, but the answers themselves do not completely overlap since they apply to the different regions of the spectrum. This follows if one remembers that for each instant of time t one must include in the dynamical equations a contribution to the density from the "waves-that-are," satisfying $\lambda < ct$, but not from the "waves-to-be," satisfying $\lambda > ct$.

We summarize the results in Table 15. In so doing, we characterize the gravitational-wave energy by its ratio to the energy of the 2.7 K electromagnetic radiation.[7] The wavelengths presented refer to the present time. In each case we give the value that seems most likely. The discussion of all of the complications has been given above. In referring to Table 15 one should remember that Ω_r/Ω_m lies within the limits of 5×10^{-5} and 5×10^{-3}.

17.7 The Effect of Gravitational Waves on the Relic Radiation

Like any stationary metric perturbation, gravitational radiation influences the distribution of the 3 K background photons. According to the general laws of optics, this influence leads to differences in the observable radiation temperature T with direction. The form of the spectrum, however, does not change. In particular, if the spectrum is of the Planck equilibrium form at recombination, then it remains of the Planck form in every direction. The Liouville theorem in combination with GTR guarantees a change in the photon flux density proportional to T^3, corresponding to the change in the photon frequency and the temperature. To describe the temperature that obtains in the presence of gravitational waves, it is therefore necessary to introduce a function $T(\theta, \phi)$ or $\Delta T(\theta, \phi) = T(\theta, \phi) - \langle T \rangle$.

The most important studies of this temperature distribution were carried

TABLE 15

The Relative Density Ω_g/Ω_r of Gravitational Radiation in the Universe as Given by Several Different Arguments

Argument	Result	Spectral Region of Applicability
Nucleosynthesis	$\Omega_g/\Omega_r < 15$	$0 < \lambda < 3 \times 10^{17}$ cm
Perturbation evolution	$\Omega_g/\Omega_r < 200$	$0 < \lambda < 5 \times 10^{26}$ cm
The age of the Universe	$\Omega_g/\Omega_r < 2 \times 10^4$	$0 < \lambda < 10^{28}$ cm

7. Note though that the Hubble parameter and the material density do not enter into the result here.

out by Dautcourt (1969) at the Potsdam Observatory in the German Democratic Republic. Without reproducing the mathematical details, we shall give a qualitative description of the results.

It is evident that the dimensionless frequency change of an electromagnetic wave along a given ray or the dimensionless temperature change along a given ray can be expressed in terms of the dimensionless gravitational wave amplitude h. (For the sake of brevity, we are not writing the indices: what we are interested in, of course, are corrections to the metric components transverse to the propagation direction.) Since h decreases in the course of the expansion, the relevant quantity is the value of h at recombination. Inasmuch as the sum total of the fluctuation-induced changes satisfies $\Delta T/T < 10^{-4}$, we obtain the estimate $h(t = t_r) < 10^{-4}$. Allowing for the expansion since recombination, this corresponds to the present bound $h_0 < 10^{-7}$. But the density of gravitational-wave energy depends on dh/dt. It is therefore necessary to explain to what wavelengths and frequencies this limit applies.

In the case of short, high-frequency waves, one cannot assume that the recombination process is instantaneous. Rather, we see simultaneously several layers on which the gravitational wave effects are not all of the same sign. One can, however, observe an undiminished effect for $\lambda_r > 0.05ct_r$, which, upon conversion to present-day scales, corresponds to $\lambda_0 > 2 \times 10^{25}$ cm. On the other hand, to ensure that we are dealing with a gravitational wave that has formed, we require $\lambda < ct_r$. Putting these bounds together, we need 2×10^{25} cm $< \lambda < 5 \times 10^{26}$ cm today. For a wavelength in this interval, we estimate the ratio $\epsilon_g/\epsilon_r = \Omega_g/\Omega_r$ to lie between 10^{-8} and 4×10^{-4}.

To see whether the greatest sensitivity is achieved for wavelengths equal to the horizon size at recombination, we must examine the situation for other spectral regions. Note that horizon-sized waves give rise to temperature anisotropies ΔT on an angular scale of the order of $ct_r(1 + z_r)/ct_0$, i.e., about 2° or 0.03 rad.

In a detailed examination of the interaction of a gravitational wave with the relic radiation, several curious features come to light that are related to the fact that the speed of a gravitational wave equals the speed of light. A ray of light (a radio wave) propagating at a small angle with respect to a single gravitational wave experiences the same phase of the wave for a long time. This intensifies the interaction. On the other hand, a gravitational wave is transverse, a characteristic that weakens its effect on a ray sliding by at a small angle. As a result, there appears close to the direction of propagation of a single gravitational wave (we shall take it to be along the z-axis, i.e., the $\theta = 0$ direction in spherical coordinates) a distinctive form for the background photon distribution: $\Delta T \approx \cos(2\phi + \alpha)$.

The temperature perturbation ΔT itself remains finite, but the gradient of

ΔT on the sphere tends to infinity[8] near the axis. If one considers the decomposition over the sphere (θ, ϕ) into Legendre polynomials, the high harmonics decrease correspondingly slowly with harmonic number for a given wavelength. They decrease according to a power law, and not exponentially. For perturbations which become gravitational waves after recombination, the sensitivity of $\Delta T/T$ to h is less. For a given h, however, the energy density is also smaller. For a wavelength $\lambda = ct_0 = 10^{28}$ cm (the radius of the horizon today), the perturbation ΔT is of quadrupole character and is called a 12-hour fluctuation. The amplitude $\Delta T/T$ is still of order h. An estimate based on observations then gives $h < 3 \times 10^{-4}$ and $\Omega_g < 10^{-7}$, respectively, so that $\Omega_g/\Omega_r < 10^{-3}$.

The situation in the intermediate region 5×10^{26} cm $< \lambda < 10^{28}$ cm is clear from an interpolation between the limiting cases. Waves with $\lambda > 10^{28}$ cm do not presently exist, in the sense that the corresponding tensor perturbations are still not waves. The temperature fluctuations are highly sensitive to waves with $\lambda \sim (1\text{–}5) \times 10^{26}$ cm because the metric distortion that occurs during the recombination period, when the interaction between the electromagnetic radiation and the matter first ceases, is observable. It is assumed in this case that the secondary ionization which is related to the formation of clusters and quasars or to the even earlier formation of shock waves takes place sufficiently late ($z < 10$–20) that the fluctuations are not smeared out by the last scattering. Finally, short-wavelength gravitational waves with $\lambda < 2 \times 10^{25}$ cm produce only small fluctuations in ΔT since these fluctuations are inevitably smeared out in the process of recombination, a process which cannot be regarded as instantaneous.[9] Any gravitational waves, however, including short waves, participate in the creation of quadrupole ΔT perturbations with an amplitude corresponding to the present-day value of h. We thus obtain for the energy density the expression

$$\Omega_g/\Omega_r = 10^{-3}(10^{28}\ \text{cm}/\lambda)^2\,, \qquad \lambda \leq 10^{28}\ \text{cm}\,. \tag{17.7.1}$$

Note that this covers the limiting case $\lambda = 10^{28}$ cm considered above.

Let us now compare equation (17.7.1) with the information in Table 15. The equation specifies Ω_g for wavelengths larger than most of the ranges pertaining to Table 15, so Table 15 and equation (17.7.1) supplement each other.

17.8 The Peculiar Motion Induced by Gravitational Waves

Is it possible to use clusters of galaxies—or separate galaxies and stars within clusters—as gravitational radiation detectors? At the time of the

8. The components of the gradient are $\partial/\partial\theta$ and $\theta^{-1}\,\partial/\partial\phi$. The infinity appears on account of the factor θ^{-1}.

9. What enters the analysis is the time for the change in the optical depth to scattering along the path of a ray approaching the observer.

great enthusiasm surrounding Weber's results, Rees (1971) posed this question. He actually gave estimates, not for stellar systems, but for free particles, particles interacting neither with one another nor with other bodies. Let us go through his analysis.

If the distance between such particles equals x, then a wave gives rise to relative motion at the speed $v = (dh/dt)x$. (In this discussion we shall omit the factors of order unity, which take account of the angle between the line of separation of the particles and the direction of the principal axis of the deformation tensor of the gravitational wave.) This speed remains constant for a time of order the period $\tau = \lambda/c$ of the gravitational wave. Rees then supposes that the velocity change produced during the following period is statistically independent with respect to magnitude and direction. In this case the squares of the velocity changes add, so Rees estimates the velocity gain after a time T (corresponding to n periods) as:

$$v^2(T) = n(dh/dt)^2x^2 = \frac{T}{\tau}(dh/dt)^2x^2\,, \qquad v = \left(\frac{dh}{dt}\right)x\left(\frac{T}{\tau}\right)^{1/2}. \quad (17.8.1)$$

Let us take the upper limit on the gravitational radiation density to be the critical density. From equation (17.2.6) this means that

$$\frac{\epsilon_g}{c^2} = G^{-1}\left(\frac{dh}{dt}\right)^2 = H^2G^{-1}\,, \qquad \frac{dh}{dt} \sim H\,. \quad (17.8.2)$$

Now Hx is the relative velocity between the two bodies due to the overall expansion of the Universe. According to Rees's formula, the induced relative velocity can be $n^{1/2}$ times larger.

One finds it hard to agree with this assertion because a spectral approach is necessary in this problem. Formula (17.8.1) tacitly assumes statistical independence of the accelerations d^2h/dt^2 over intervals of time greater than τ. But statistical independence signifies a flat spectrum for the Fourier decomposition of $d^2h(t)/dt^2$ at a given point. Hence $h^{(2)}(\omega) = \text{const.}$ for $\omega < \tau^{-1}$, where $h^{(2)}(\omega)$ is the Fourier transform of d^2h/dt^2. But in this case $h^{(1)}(\omega) = \text{const.}/\omega$, so the energy density, equal to $\int[h^{(1)}(\omega)]^2d\omega$, diverges as $\omega \to 0$! The quantity $h(\omega) = \text{const.}/\omega^2$ diverges even more strongly. The assumption of independent velocity gains over periods of duration τ therefore turns out to be far from straightforward.

The properties of a statistically independent collection of gravitational waves with a flat power spectrum, $|h^{(1)}(\omega)|^2 = \text{const.}$, on the other hand, are very different. The acceleration is variable in sign, so its distribution is anticorrelated. The velocity therefore remains constant and equals $\langle(dh/dt)^2\rangle^{1/2}x = (\epsilon_gG/c^2)^{1/2}x$. The shift relative to the unperturbed location grows like $t^{1/2}$.

What do observations say about this? According to Sandage's data, the deviations of observed redshifts from the mean curve are of order 10% over a wide range of distances from 10 Mpc to 2000 Mpc. That the absolute

value of the Hubble parameter is only known to an accuracy of order 100% does not contradict this assertion because this indeterminancy relates only to the overall scale factor. However, only a part of the deviation on the (log z, log M)-curve in the data for individual clusters depends on the peculiar contribution to the relative motion. Therefore, conclusions about the gravitational wave density that are based on Sandage's data only constitute an inequality. The data only imply an upper limit. Hence, for wavelengths in the interval from 3×10^{25} cm to 6×10^{27} cm, we obtain the estimate

$$\left(\frac{dh}{dt}\right)^2 < 0.01H^2\,, \quad \Omega_g < 0.01\,, \quad \frac{\Omega_g}{\Omega_m} < 100\,. \tag{17.8.3}$$

We now return to the theoretical side of the question.

Note that one cannot get information from the known small velocity of our Galaxy (or of the Local Group of galaxies) relative to the relic radiation. The point is that a gravitational wave causes both the matter (the cluster) and the gas (the radiation) surrounding it to move in the same way. There is no *local* relative velocity.[10] Measurements of the relative velocities of distant galaxies (i.e., of the Hubble parameter) cannot therefore be replaced by local measurements. Are there also circumstances in which the kinetic energy of relative motion grows monotonically in a gravitational-radiation field? The moral of the preceding discussion is that (dh/dt) changes sign over the course of time and the average value of $\int (dh/dt)dt$ is smaller than it would be for a "random" h.

A monotonic energy gain is possible, however, if the relative velocity also changes its sign[11] when the motion is unperturbed by a gravitational wave. A simple example where this occurs is a gravitationally bound system of bodies (a cluster) in which each body moves along a more or less periodic orbit.[12] In this case the energy gain is proportional to the spectral density of the gravitational radiation in the corresponding frequency region. But then the aggregate of particles also radiates gravitational waves. Whether energy is gained or lost depends on the relation between the effective temperature of the radiation and the particles. Let us consider this point in more detail.

Suppose that particles with mass m and velocity u move along a trajectory with radius R. The kinetic energy and the temperature of a particle are then of order mu^2. The frequency of the emitted gravitational wave is of order u/R, and the corresponding wavelength is $\lambda = cR/u$. According to the Rayleigh-Jeans formula, thermal radiation at a wave-

10. Only more subtle local effects, related to the collisionless nature of the gas, remain. These include, for example, the quadrupole deviations in $\Delta T(\theta, \phi)$ discussed in § 17.7.

11. What is meant is the velocity component along a specified axis.

12. For Newtonian motion, exact periodic motion takes place in the field of a central mass or inside a region in which the density is constant.

length of order λ (viz., in the wavelength range λ to ∞ or $\lambda c^{-1/2}$ to $\lambda c^{1/2}$) has an energy density of order $\epsilon_{g,\mathrm{eq}} = T\lambda^{-3}$; this gives $\epsilon_{g,\mathrm{eq}} = mu^5R^{-3}c^{-3}$.

The particles gain energy providing the gravitational wave energy density is small compared to the binding energy density of the particles moving along the trajectory. According to the virial theorem the binding energy of the particles is of order their kinetic energy. Using this relation we estimate the characteristic time τ required for the dissociation of the aggregate of particles as the time required for an energy gain equal to the initial kinetic energy.[13] Now, for $\epsilon_g \gg \epsilon_{g,\mathrm{eq}}$, the rate of energy gain is

$$\left\langle \frac{d}{dt}\left(\frac{mu^2}{2}\right)\right\rangle = \frac{GmuR\epsilon_g}{c^2} \tag{17.8.4}$$

for a wide range of gravitational wave spectra. The characteristic time τ is thus

$$\tau \equiv \left(\frac{mu^2}{2}\right)\Big/\left\langle \frac{d}{dt}\left(\frac{mu^2}{2}\right)\right\rangle = \frac{uc^2}{GR\epsilon_g} = \frac{t_0^2}{\Omega_g \tau_0}, \tag{17.8.5}$$

where $\tau_0 = R/u$ is the period of the unperturbed particle motion. The gravitational radiation energy density ϵ_g can be expressed in terms of the parameter $\Omega_g = \rho_g/\rho_c = \epsilon_g/(c^2\rho_c)$. Here ρ_c is given in terms of the age of the Universe t_0, to an order of magnitude, as $\rho_c = (Gt_0^2)^{-1}$.

We know for certain that $\Omega_g \leq 1$ (see the discussion in § 17.6 concerning the age of the Universe). Furthermore, $\tau_0 < t_0$; this is a direct consequence of the fact that the system is isolated and gravitationally bound. But equation (17.8.5) then implies that $\tau > t_0$. In other words, the gain in energy over the entire cosmological expansion time is small.

17.9 The Interconversion of Gravitational and Electromagnetic Waves

The processes of wave interconversion are interesting in several respects. First, it is in principle possible that such processes take place in the Universe. Second, these processes are discussed as a method for the laboratory generation of high-frequency, short-wavelength gravitational waves. Finally, the detection of gravitational waves requires their prior conversion into other forms of energy. For short waves, of cosmological or laboratory origin in particular, it is natural to seek a method for converting them into electromagnetic waves.

A most important case is the conversion of waves in the presence of a magnetic field H_0 that is constant (independent of the space and time coordinates) and transverse with respect to the propagation direction. We shall assume here that the field is oriented along the y-axis and that the wave is propagating in the direction of the x-axis. A peculiar feature of the

13. We are here not accounting for the change in the parameters u and R that occurs during the dissociation.

phenomenon, a feature related to the equality of propagation speeds of electromagnetic and gravitational waves, was first noticed in a remarkable study by Gertsenshteyn (1961). A plane electromagnetic wave in empty space does not emit gravitational waves since the stress-energy tensor of such a wave does not contain a variable quadrupole moment. As is well known, only a stress-energy distribution with such a moment gives rise to the emission of gravitational waves. However, a time-variable quadrupole term does appear when an electromagnetic wave propagates through a constant magnetic or electric field. We shall now study this effect.

An electromagnetic wave with amplitude H_y in a field H_0 produces a quadrupole stress term in the Maxwell tensor $T_{yy} = H_y H_0 \cos(kx - \omega t)$. This stress tensor then enters on the right side of the equation for the radiation of gravitational waves (see Landau and Lifshitz 1962):

$$\Box h_{yy} = GT_{yy} = GH_y H_0 \cos(kx - \omega t) . \tag{17.9.1}$$

Inasmuch as $\cos(kx - \omega t)$ is a particular solution of the wave equation, i.e., it gives a resonance in the presence of such a term on the right-hand side, the solution takes the form

$$h_{yy} = ax \cos(kx - \omega t) . \tag{17.9.2}$$

The constant field turns on at the origin of the coordinates:

$$H = 0 , \quad x < 0 ; \quad H = H_0 , \quad x > 0 . \tag{17.9.3}$$

This means that the fraction of electromagnetic-wave energy that is converted into gravitational-wave energy is[14]

$$\alpha = GH_0^2 x^2 / c^4 . \tag{17.9.4}$$

Many authors point out that coherence is also important for the reverse process. Thus, upon propagating through a field strength H_0, a gravitational wave gives rise to magnetic field perturbations that act as sources in the equation

$$\Box H_y = \omega^2 h_{yy} H_0 = \omega^2 H_0 h \cos(kx - \omega t) \tag{17.9.5}$$

for the electromagnetic waves. The analogy with the preceding case is complete; the energy conversion coefficient for this, the reverse process, is also α.

In the laboratory and even for pulsars α is small. If $H_0 = 10^5$ gauss and $x = 10^2$ cm, we find $\alpha = 10^{-35}$; if $H_0 = 10^{13}$ gauss and $x = 10^6$ cm, $\alpha = 10^{-11}$.

14. In this context we should correct an erroneous assertion contained in Volume 1 where, instead of x^2 in the expression for α, there appeared x. We were then solving the problem of the conversion of a spherical electromagnetic wave into a plane gravitational wave. In reality, a spherical electromagnetic wave is converted into a gravitational wave that is also spherical. (We are assuming that $x \gg \lambda$ and that the spherical angle is less than 4π.) In this case the conversion coefficient is $\sim x^2$, in agreement with equation (17.9.4).

For a cosmological field one could make the extreme assumption that $H_0^2 = \epsilon_r$ ($H_0 = 10^{-6}$ gauss today and $H_0 = 1$ gauss at recombination). One would then obtain $\alpha \sim 10^{-3}$ at $z = z_r$. Such a conversion could be noticed in the relic radiation by the characteristic weakening of a single polarization of the radio waves that would result. If, however, the conversion occurs in a medium consisting of atoms and electrons whose conductivity is nonzero and whose dielectric permeability differs from unity, the coherence is destroyed. It then turns out that α remains immeasurably small under cosmological conditions.

Notice here the symmetry between equations (17.9.1) and (17.9.5). To make this more explicit, we use the notation $f = H_y$, $p = H_0\sqrt{G}$, and $g = (dh/dt)/\sqrt{G}$, to obtain the alternative equations

$$\Box f = \omega p g\,, \qquad \Box g = \omega p f\,. \tag{17.9.6}$$

The normal modes for such a system of equations are mixed (gravitational-electromagnetic waves). If the phases are the same, the solution is denoted by q; if they are opposite, the solution is denoted by r. For such modes, the equations separate to give

$$\Box q = \omega p q\,, \qquad \Box r = -\omega p r\,. \tag{17.9.7}$$

One can thus construct an elegant, exact theory of interacting waves.[15] However, it is not apparent where such a situation is realized in nature.

If electrons and atoms are present, the equation for an electromagnetic wave takes the form

$$\Box H_y = \omega^2 h_{yy} H_0 + (a + ib) H_y\,, \tag{17.9.8}$$

where a and b are the real and imaginary parts of the polarization of the medium. The symmetry between the two equations is then destroyed and, as noted above, the effects become negligibly small even under cosmological conditions.

The problem of the electromagnetic detection of gravitational waves is similar to the preceding problem. Imagine a resonator consisting of a closed metal container with walls having the highest possible conductivity. The particular solutions of the Maxwell equations for the fields in this container, which satisfy the boundary conditions at the walls, take the form

$$E,H = \phi_k(t) f_k(x)\,. \tag{17.9.9}$$

In the absence of a gravitational wave and of loss terms, the equation for the function ϕ_k and its solution are

$$\frac{d^2\phi_k}{dt^2} + \omega_k^2 \phi_k = 0\,, \qquad \phi_k = c_k \exp(-i\omega_k t)\,. \tag{17.9.10}$$

15. The coefficient α describes the initial stage when one wave is weak and the other is just beginning to appear.

Now suppose that such a resonator is located in a region of constant magnetic field and subject to the action of a gravitational wave with frequency ω. Then there appears on the right-hand side of the equation for ϕ_k a forcing term:

$$\frac{d^2\phi_k}{dt^2} + \omega_k^2\phi_k = \omega^2 H_0 h e^{-i\omega t} . \tag{17.9.11}$$

The best situation for detection occurs when $\omega = \omega_k$. In this case, given the initial value $\phi_k = 0$, the solution takes the form

$$\phi_k = \text{const.} \times t \exp(-i\omega_k t) . \tag{17.9.12}$$

Now the energy of the resonant oscillation that accumulates after a time t is proportional to t^2. If one relates this to the gravitational wave energy that passes through the resonator during this time (proportional to t), one obtains the "utilization coefficient"

$$\alpha = GH_0^2 xct/c^4 . \tag{17.9.13}$$

It is evident that the maximum value of t is determined by the decay of the oscillations in the resonator and the characteristic decay time τ can indeed be large! The utilization coefficient of the resonator is $c\tau/x \approx \omega_k\tau = Q$ times greater than the conversion factor of gravitational waves into electromagnetic waves in open space. Here Q is the so-called figure of merit, the number of oscillations of the resonator which are possible before appreciable decay.

A modification of this method is the following. Assume that the mode n with frequency ω_n is already excited in the resonator and that the action of a gravitational wave excites, say, the k mode with the frequency ω_k. Resonance is then achieved for gravitational-wave frequencies $\omega = \omega_k - \omega_n$ or $\omega = \omega_k + \omega_n$.

The excited mode k can have a nonzero initial amplitude $A_{k,0}$. After the action of the gravitational wave, the newly acquired amplitude gain δA_k (proportional to $\omega_k^2 ht$) combines with the initial amplitude $A_{k,0}$. For an optimal choice of phase, the energy gained after a given time turns out to be larger than it would have been previously: $\Delta E_k \sim ht$, not $\Delta E_k \sim h^2t^2$.

However, it is clearly more difficult to measure an energy change ΔE_k on a background of initially nonzero energy than it is to measure a change from zero energy.

One can also excite a particular mode m in the resonator and study the change in the character of this mode under the action of a gravitational wave with twice the frequency, $\omega = 2\omega_m$. (This is a particular case of the previous formula with $\omega = \omega_k + \omega_m$ for $\omega_k = \omega_m$.) Such a situation is called parametric resonance. Under the action of the gravitational wave, the amplitude and phase of the electromagnetic oscillation in the resonator

slowly change. The governing equations are

$$H_k = A_k(t) \cos [\omega_k t + \psi_k(t)] ,$$

$$\frac{dA_k}{dt} = A_k \left|\frac{dh}{dt}\right| \cos 2\psi_0 , \quad \frac{d\psi_k}{dt} = \left|\frac{dh}{dt}\right| \sin 2\psi_0 . \quad (17.9.14)$$

Here ψ_0 is the phase shift of the gravitational wave relative to the electromagnetic oscillation. In its dependence on the choice of ψ_0, the electromagnetic oscillation can be intensified, weakened, or phase-shifted under the action of the gravitational wave. The effect is proportional to dh/dt, i.e., to $\epsilon_g^{1/2}$.

The use of electromagnetic detectors is apparently the most promising method for laboratory experiments concerned with the detection of high-frequency gravitational waves that are themselves generated in the laboratory with known frequency and phase.

However, even for cumbersome devices with fields of 10^5 gauss and with sizes of tens of meters, the estimates still do not give acceptable results. Indeed there is no extant project which employs this method.

On the other hand, the theoretical advantage of a resonator would be lost in the detection of a wide-band spectrum of cosmological gravitational radiation. This is because its sensitivity is large only in a narrow frequency band, and the sensitivity averaged over the spectrum is no better than for a magnetic field in open space. For the laboratory conditions $H = 10^5$ gauss and $x = 100$ cm given above, we find that the coefficient describing the conversion of gravitational waves into electromagnetic waves is of the order of 10^{-35}. If we take the number density of cosmological gravitons to be 10^3 cm^{-3}, their flow through the area $x^2 = 10^4$ cm^2 amounts to 10^{17} s^{-1}. Hence, under the influence of such radiation, the probability of photon creation over the entire volume is $10^{17} \times 10^{-35}$ s^{-1} $= 10^{-18}$ s^{-1}. This amounts to the creation of one photon in the age of the Universe. Apparently, therefore, the laboratory observation of low-frequency gravitational waves from binary stars and pulsars is far closer to realization than is the observation of high-frequency, relic, gravitational radiation.

IV ANISOTROPIC COSMOLOGY

18 INTRODUCTION TO PART IV

There is every reason for supposing that at present and in the recent past the Metagalaxy obeys the simplest and most "natural" assumptions concerning the isotropy of motion and the homogeneity of the distribution of the matter contained within it. We are here diverting our attention from those small-scale (by comparison with the whole observable Metagalaxy) inhomogeneities that manifest themselves as galaxies and clusters of galaxies. As was already noted, a decisive argument in favor of isotropy and homogeneity is the isotropy of RR in the hot Universe. (The upper limit on possible anisotropy of the radiation is $\Delta T/T \leq 0.003$.)

The isotropy of the radiation testifies to the identity of the conditions in different directions. Since the background radiation in the well-studied region of the spectrum from $\lambda = 0.25$ cm to $\lambda = 20$ cm interacts weakly with dust, with neutral atoms, and with plasma, this permits one to draw conclusions concerning the isotropy at larger distances from us than the distances about which one can draw conclusions by means of the statistics of distant discrete objects. Indeed, the photons incident on us today scatter at a distance that corresponds on average to a redshift of not less than $z = 6$ or $z = 8$, and perhaps at $z = 1400$. Hence, as will be shown below, observations demonstrate that the expansion of the Universe is quite isotropic at least as early as $t \approx 0.01 t_0$, where t_0 represents the present time. It is more probable, though, that the expansion is isotropic from as early as $t \approx 10^{-4} t_0$.

On the one hand, therefore, it is an observational fact that the Universe is at present isotropic and homogeneous. On the other hand, as Misner (1969*a*) particularly has emphasized, this very fact of the Universe's isotropy and homogeneity is puzzling.

Actually, if the Universe becomes transparent at $z = 7$, then photons incident on us from directions separated on the sky by more than 30° originate at points which, at the moment of emission, are separated by a distance greater than the optical horizon. In other words, there is no causal connection between such photons. Thus, the temperature at either one of these points cannot be known at the other point. There cannot be any mechanism for equalization of the temperature and the smoothing of any inhomogeneities. It is therefore necessary to assume that strict isotropy and homogeneity are imposed at the singularity.

On the other hand, the assumption of strict homogeneity and isotropy for the singular state can in no way be considered provable from any general principles. Observational data simply do not yield unambiguous information concerning the singularity.

The two following questions therefore arise:

1. Is it really true that the Universe expands homogeneously and isotropically from the very beginning, or at least from the time when the most general laws of physics known today apply, i.e., when the curvature of spacetime is $\sim(10^{-33}\,\text{cm})^{-2}$? If so, what are the reasons for the strict homogeneity and isotropy of the singular state?

2. If the initial stages of expansion exhibit non-Friedmannian behavior, then just what are these stages like and why does the Universe finally become homogeneous and isotropic?

To answer these questions, it is necessary to examine models whose initial stages exhibit non-Friedmannian behavior and to determine those observable properties from which one can determine the properties of the early stages of expansion. In addition, if the beginning of the expansion is actually non-Friedmannian, then one must elucidate those causes or those processes which lead to the present-day Friedmann picture.

If one could successfully show that for highly arbitrary initial conditions the Universe must emerge, in the course of the expansion, as a Friedmann solution, then one could say that the homogeneity and isotropy are explained. This is why research on more general solutions has very recently followed on the heels of the elucidation of all of the properties of the very simple Friedmann model.

We must stipulate here, however, that the aggregate of all of the observational and theoretical information forces us to tend toward the first of the possibilities mentioned above. This is, namely, the possibility that the cosmological expansion is homogeneous and isotropic from that time when classical (nonquantum) GTR is applicable to a description of the Universe, i.e., from $t_{\text{Pl}} = 10^{-43}$ s. The possible reasons for the homogeneity and isotropy of the singular state will be considered in the next part of the book.

We shall also, however, consider the possibility that the early stages of expansion exhibit substantially non-Friedmannian behavior. As was noted above, such a possibility cannot presently be excluded. This part of the book is therefore devoted to an examination of the particular case of strictly homogeneous but anisotropic solutions of the GTR equations. This class of solutions is distinguished not only by its relative mathematical simplicity, which in many cases permits one to study an exact solution, but also by the hope that homogeneous anisotropic (HA) cosmological solutions reflect at least several important features of the general case (Lifshitz and Khalatnikov 1969; Belinsky and Khalatnikov 1969*a*, *b*; Belinsky, Lifshitz, and Khalatnikov 1970; Zel'dovich 1969; Misner 1969*a*, *b*).

As Misner (1969*a*) has shown, in a definite class of HA models with a closed, three-dimensional space, light can circumnavigate the Universe repeatedly during an early stage. Consequently the difficulty mentioned above concerning the presence of an optical horizon does not exist. These models therefore admit—whereas the Friedmann model does not—the theoretical possibility that inhomogeneities equalize over a large scale. It is true, though, as was explained later (Doroshkevich, Lukash, and Novikov 1971), that the class of such models is very narrow.

There is additional interest in HA models because the absence of isotropy opens up the possibility of considering a primordial magnetic field or a directed flux of matter or free particles. Such a flux *can* in particular lead to nonequilibrium processes, and increase in entropy, and so forth.

It is possible that one will be able to obtain the inhomogeneous case by the superposition of "perturbations" on a homogeneous model (Zel'dovich 1969; Khalatnikov 1965).

All this, along with the possibility of a tractable mathematical analysis of the solution, makes the examination of the HA models quite important.

We shall begin with a consideration of the kinematics of the expansion of HA models. As in the case of the Friedmann solution, we shall start with a Newtonian problem (Lynden-Bell 1962, 1964; Zel'dovich 1964*b*), and then pass to a GTR analysis (Schücking and Heckmann 1958; also work done subsequently by many other authors). We emphasize immediately that the Newtonian and relativistic analyses of the HA problem in many respects give results that are not similar, in distinction to what happens in the study of the homogeneous and isotropic case.

After an examination of the basic properties of the kinematics of HA models and of the growth of perturbations in them, we shall consider the characteristics of the physical processes taking place in them and *possible* observational predictions. The last chapters in this part of the book are devoted to a general analysis of homogeneous models and to a comparison of them with observations.

19 VERY SIMPLE ANISOTROPIC COSMOLOGICAL MODELS

19.1 The Newtonian Theory of a Very Simple Homogeneous, Anisotropic (HA) Model as the Limiting Case of a Local Problem

We have previously constructed the cosmological solution in the isotropic and homogeneous case by beginning with an examination of the expansion of a finite sphere. We shall also try to construct an HA cosmological solution by proceeding from an analysis of a problem for a finite body—a triaxial ellipsoid filled with matter of constant density. As was done in § 1.1, we shall here consider pressure-free ($P = 0$) matter.

It is well known (see, for example, Volume 1) that in Newtonian theory the addition of new layers to an already existing ellipsoidal distribution of homogeneous matter without a change in shape does not change the gravitational field inside the original distribution. Thus, suppose that one manages to find a solution for the motion of the matter constituting a homogeneous ellipsoid such that the motion always has the effect of converting the original ellipsoid into another homogeneous ellipsoid, though the latter may have a shape, a spatial orientation, and a scale that are different from those of the former. Then, by adding an unlimited number of new layers that preserve the shape of the ellipsoid (a process that in no way affects the motion of the matter composing the inner parts of the ellipsoid), we obtain a homogeneous cosmological solution that is anisotropic, or, generally speaking, deformed.

Let us now consider the homogeneous matter filling a three-dimensional ellipsoid. As is well known (see, for example, Sretenskiy 1946), the gravitational field inside an ellipsoid is a quadratic function of the coordinates:

$$\phi = \phi_0 + \tfrac{1}{2}\phi_{ik}x_i x_k \,. \tag{19.1.1}$$

The coefficients ϕ_0 and ϕ_{ik} do not depend on coordinates. The axes of the tensor ϕ_{ik} coincide with the axes of the ellipsoid. If one chooses the coordinate axes along the axes of the ellipsoid, so that $x^2/a^2 + y^2/b^2 + z^2/a^2 = 1$ is the equation of the surface of the ellipsoid, then, inside it,

$$\phi = \phi_0 + \tfrac{1}{2}\phi_{xx}x^2 + \tfrac{1}{2}\phi_{yy}y^2 + \tfrac{1}{2}\phi_{zz}z^2 \,. \tag{19.1.2}$$

In accordance with Poisson's equation, we then have

$$\phi_{xx} + \phi_{yy} + \phi_{zz} = 4\pi G\rho \,. \tag{19.1.3}$$

The functions ϕ_{xx}, ϕ_{yy}, and ϕ_{zz} are not equal to one another; their relation depends on the relation among the ellipsoid axes. The largest value ϕ_{ii} (here there is no summation over the indices ii!) with the greatest magnitude is the one for the direction x_i corresponding to the shortest ellipsoid axis. For example, in the limit $a = b \ll c$ of a prolate ellipsoid stretched out along the z-axis,

$$\frac{\partial^2\phi}{\partial x^2} = \frac{\partial^2\phi}{\partial y^2} = 2\pi G\rho\,, \qquad \frac{\partial^2\phi}{\partial z^2} \to 0\,. \tag{19.1.4}$$

For a thin disk (a highly oblate spheroid) with $a = b \gg c$,

$$\frac{\partial^2\phi}{\partial z^2} = 4\pi G\rho\,, \qquad \frac{\partial^2\phi}{\partial x^2} = \frac{\partial^2\phi}{\partial y^2} \to 0\,. \tag{19.1.5}$$

The correct formulation of the problem in Newtonian theory goes as follows. We specify initially the density ρ and a generalized Hubble-like linear velocity field (Narlikar 1963)

$$u_i = H_{ik}(t_0)x_k\,. \tag{19.1.6}$$

As we shall see later, this specification guarantees that the necessary properties of the solution are maintained. Formula (19.1.6) allows for the presence of rotation along with deformation if $H_{ik} \neq H_{ki}$. We shall also specify the equation[1] $A_{ik}(t_0)x_ix_k = 1$ for the ellipsoid's surface, the surface that separates the region filled by matter (the inside) from the vacuum.

The choice of an ellipsoidal shape for the shape of the matter-filled region guarantees a quadrupole form for the potential ϕ. This then leads to a linear dependence of the acceleration on coordinates:

$$\frac{d^2x_i}{dt^2} = -\phi_{ik}x_k\,. \tag{19.1.7}$$

It follows from this that the dependence of the velocity on the coordinates remains linear in the course of time. Therefore, the components of the tensor H_{ik} depend on time but not on the coordinates. Now the equation for H_{ik} takes the form

$$\frac{dH_{ik}}{dt} = -H_{in}H_{nk} - \phi_{ik}\,. \tag{19.1.8}$$

In turn, the linear coordinate dependence of the velocity guarantees that homogeneity is preserved with time, since $\rho^{-1}\,d\rho/dt = -\nabla\cdot\boldsymbol{u} = H_{ll}$ does

1. It is convenient to write the equation of the ellipsoid in a form that allows for the possible departure of the ellipsoid axes from the coordinate axes over the course of time. The quantities ϕ_0 and ϕ_{ik} are found by joining the internal solution (19.1.2) to the solution outside the ellipsoid. The latter solution satisfies the conditions on $\nabla^2\phi$ and the condition that $\phi = 0$ at infinity.

not depend on the coordinates (here, a summation is carried out over the index l). Moreover, since the velocity depends linearly on the coordinates, the shape of the surface bounding the matter always remains ellipsoidal even though the coefficients A_{ik} change with time according to the equation

$$\frac{dA_{ik}}{dt} = -H_{li}A_{lk} - H_{mk}A_{im}\,. \tag{19.1.9}$$

It should be emphasized that it is not only the spatial orientation of the axes that changes but also the shape of the ellipsoid as characterized by the ratios of its semiaxes. So, in the particular case when the matter is initially at rest, the density goes to infinity after a certain amount of time. This is because the initially shortest axis goes to zero; the two other axes, meanwhile, decrease but remain finite, so that at the instant of infinite density the ellipsoid degenerates to a plane. Now let us assume, on the other hand, that the matter initially fills a sphere, so that the tensor ϕ_{ik} is *initially* a unit or isotropic tensor. Then, $\phi = \phi_0 + 2\pi G\rho r^2/3$. Suppose, however, that the initial velocity distribution is anisotropic, that the tensor H_{ik} is not a unit tensor. Then, in the course of time, the sphere deforms, transforming into an ellipsoid. The initial isotropy of the ϕ_{ik} is thus destroyed.

The system of equations for dH/dt and dA/dt become a closed system upon a specification of the dependence of ϕ_{ik} on A_{ik} and ρ, a dependence which is given by potential theory (see Sretenskiy 1946). As is well known, this dependence can be expressed in terms of elliptic integrals. From the resulting expressions, it is easy to see that ϕ_{ik} is symmetric and positive definite.

We shall not analyze all of the properties of the solution of the problem, so we refer interested readers to Zel'dovich's (1964*b*) study. However, we shall discuss the case that is interesting from the point of view of applications to the cosmological problem of deriving an HA solution.

Thus, consider first the case in which rotation is absent. Then H_{ik} is symmetric, and the principal axes of the tensor H_{ik} coincide with the ellipsoid axes. Moreover, we shall from the outset confine ourselves to the case $a = b$ of axial symmetry (an ellipsoid of rotation or *spheroid*). We furthermore shall require that for very large scales (correspondingly, for $\rho \to 0$) the shape of the ellipsoid approach the shape of a sphere over the course of expansion. In other words, "isotropization" must ensue, and the expansion rate at infinity must tend to zero. The latter is of course equivalent to specifying the case of critical density.

The simplest procedure is not to examine the problem of anisotropic expansion but rather to examine the reverse problem of contraction and then perform a time inversion. Consider, therefore, a large sphere that is weakly deformed into an ellipsoid. Two cases are possible: $a = b > c$, where the ellipsoid is flattened like a pancake; and $a = b < c$, where the

ellipsoid is stretched out like a cigar. In both cases, $|a - c| \ll a$ holds at first.

Let us now analyze the first case, where $a = b > c$. As was noted above, the acceleration along the shortest axis is greater, because $\phi_{xx} = \phi_{yy} < \phi_{zz}$. As time increases, the flattening grows (see Fig. 53) and at some instant t_∞ the body transforms into a flat pancake satisfying $c = 0$. The bulk density of matter at this instant is infinite. The gravitational energy of the pancake is negative and finite. In spite of the infinite value of the density, the gravitational field is also finite at any point. From this it is apparent that the velocity too remains finite. Consequently, in the asymptotic region $\rho \to \infty$, the laws governing the change with time of the quantities c and ρ are

$$c \sim (t_\infty - t) \; ; \qquad \rho \sim (t_\infty - t)^{-1} \, . \tag{19.1.10}$$

Consider next the second case, in which $a = b > c$ and $\phi_{xx} = \phi_{yy} > \phi_{zz}$. In the course of time, the ellipsoid becomes still narrower (see Fig. 54) and at a certain time $t = t_\infty$ it transforms into a segment with $a = b = 0$ and c finite. The force near a long thread grows without bound as r^{-1}, where r is the distance from the thread. Consequently, the velocity in the Newtonian theory tends to infinity:

$$\left(\frac{da}{dt}\right)^2 \sim \ln \left(\frac{a_0}{a}\right)^2 , \tag{19.1.11}$$

where a_0 is of the order of $c(t_0)$.

Let us now pass to a consideration of the general case wherein the rate of contraction along all three axes is different. From the preceding account

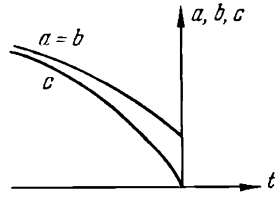


FIG. 53.—The collapse of an oblate ellipsoid to a disk. The graph displays the time evolution of the ellipsoid's principal axes a, b, and c, with $a = b > c$.

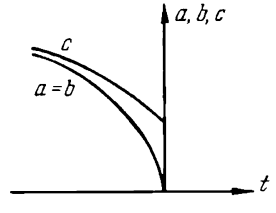


FIG. 54.—The collapse of a prolate ellipsoid to a line. The graph displays the time evolution of the ellipsoid's principal axes a, b, and c, with $a = b < c$.

the character of the solution is clear. The acceleration along the shortest axis is greatest and the body transforms at some moment t_∞ into a flat ellipse with $a \neq b$ (but not a circle, as is the case when $a = b$) and $c = 0$. It is evident that the asymptotic behavior is given by the law (19.1.10) here too.

Thus, the case wherein the ellipsoid stretches out into a thread is degenerate in the general Newtonian problem, for it requires that the two quantities a and b simultaneously go to zero. The presence of rotation in a generalized Hubble velocity distribution, i.e., the case wherein $H_{ik} \neq H_{ki}$, does not change the result qualitatively and leads merely to a rotation of the ellipsoid axes in space.

It now remains to reverse the flow of time (to consider expansion) and to increase without bound the axis lengths of the ellipsoid at a fixed instant while maintaining a fixed density. This increase is accomplished by adding new layers of matter. As a consequence of the homogeneity, the quantities a, b, and c characterize the change in the distance between any pairs of points along the x-, y-, and z-axes, respectively. But by virtue of what was said at the beginning of this section, the transition from a finite ellipsoid to an unbounded distribution in no way changes the dependences of a, b, and c on time.

We have thus constructed an HA cosmological model in Newtonian theory. For the analysis of more complicated models see Shikin (1970).

19.2 A Paradox Involving Newtonian Gravitational Theory

Could we have solved the problem for an unbounded expansion by using from the outset the equations of motion and the Poisson equation $\nabla^2\phi = 4\pi G\rho$ without first considering the intermediate problem of ellipsoids? Would this not be simpler? Notice first of all that the potential ϕ and the gravitational field vector $\nabla\phi$ are unobservable quantities. The observable quantities are the second derivatives $\phi_{ik} = \partial^2\phi/\partial x_i \partial x_k$ on which the relative acceleration of neighboring particles depends. Only one condition is imposed upon the ϕ_{ik}—that they satisfy the Poisson equation. Consequently, five degrees of freedom remain. (Keep in mind that $\phi_{ik} = \phi_{ki}$.) Thus, the equations of mechanics and the Poisson equation are not enough to determine a solution of the cosmological problem. One ought to call this arbitrariness in the choice of the ϕ_{ik} in the Newtonian theory for the case of an infinite distribution of homogeneous matter, a gravitational paradox. The common gravitational paradox is the divergence in ϕ or $\nabla\phi$ for an infinite matter distribution. However, inasmuch as ϕ and $\nabla\phi$ are unobservable, the difficulties related to the limits $|\phi| \to \infty$ and $|\nabla\phi| \to \infty$ do not arise,[2] and one ought not to call this situation a paradox.

2. In this sense, the statements by Landau and Lifshitz (1973) about the difficulties of Newtonian theory as applied to infinite space are not quite correct.

Many authors have considered the question of the elimination of the indeterminancy in the ϕ_{ik} in Newtonian cosmology. Narlikar (1963), for example, suggested the requirement that the quantities ϕ_{ik} be isotropic given an anisotropic H_{ik}. A second method consists of a preliminary examination of a finite body followed by a transition to an infinite limiting case. This we dealt with in the preceding section, taking the ellipsoid as a finite body. Clearly, however, any method for choosing a supplementary condition for the ϕ_{ik} involves a step beyond the limits of strict Newtonian theory.

Evidently, this ambiguity in Newtonian theory exists simply as a result of the examination of an infinite space filled with matter. In a problem that involves a density that falls sufficiently rapidly at infinity, one can establish the condition $\phi = 0$ at infinity. Then this condition, together with the Poisson equation, completely determines the potential.

Zel'manov (1959*a*) proposes as supplementary conditions for the removal of the ambiguity that one take those following from the analogous problem in relativistic cosmology. This evidently assumes that the solution of the relativistic problem is already known. Nevertheless, such an approach turns out useful in an examination of local perturbations in certain models, for example. This is because the approach permits one to apply the incomparably simpler apparatus of Newtonian theory. Indeed, we shall return to this point after we examine the relativistic problem.

In conclusion, we emphasize once again that the Newtonian theory does not yield a closed solution of the problem. In completely closed form, the problem can only be solved in terms of GTR.

19.3 A Simple Relativistic Model: the "Vacuum" Solution near a Singularity

We now begin a study of an HA solution of the GTR equation. The solution will be rotation-free but will have a Euclidean comoving space. In this case, evidently, all of the points of the three-dimensional space are equivalent. That is, the solution is actually homogeneous although the expansion rate in different directions can be different.

The metric for such a solution can be written in the form

$$ds^2 = dt^2 - a^2(t)dx_1^2 - b^2(t)dx_2^2 - c^2(t)dx_3^2 , \qquad (19.3.1)$$

where the speed of light is set equal to unity. The functions a, b, and c depend only on time. If these functions are not all the same, then the expansion is anisotropic. The Einstein equations for this metric are as follows:

$$a^{-1}\frac{d^2a}{dt^2} + a^{-1}\frac{da}{dt}\left(b^{-1}\frac{db}{dt} + c^{-1}\frac{dc}{dt}\right) = -8\pi G(T_1^1 - \tfrac{1}{2}T) ; \qquad (19.3.2)$$

$$b^{-1}\frac{d^2b}{dt^2} + b^{-1}\frac{db}{dt}\left(a^{-1}\frac{da}{dt} + c^{-1}\frac{dc}{dt}\right) = -8\pi G(T_2^2 - \tfrac{1}{2}T) ; \qquad (19.3.3)$$

$$c^{-1}\frac{d^2c}{dt^2} + c^{-1}\frac{dc}{dt}\left(a^{-1}\frac{da}{dt} + b^{-1}\frac{db}{dt}\right) = -8\pi G(T_3{}^3 - \tfrac{1}{2}T)\,; \quad (19.3.4)$$

$$a^{-1}\frac{d^2a}{dt^2} + b^{-1}\frac{d^2b}{dt^2} + c^{-1}\frac{d^2c}{dt^2} = -8\pi G(T_0{}^0 - \tfrac{1}{2}T)\,. \quad (19.3.5)$$

We shall now carry out an analysis of these equations.

Assume that at some instant of time the right-hand sides of equations (19.3.2–5) are much less than the component terms on the left-hand sides. Then we can examine these equations with the right-hand sides absent, in which case the solution that applies is the empty-space solution. Such a solution was obtained by Kasner (1921) and is

$$a = a_0 t^{p_1}\,; \quad b = b_0 t^{p_2}\,; \quad c = c_0 t^{p_3}\,: \quad (19.3.6)$$

$$p_1 + p_2 + p_3 = p_1{}^2 + p_2{}^2 + p_3{}^2 = 1\,. \quad (19.3.7)$$

When the relations (19.3.7) are satisfied, they leave only one arbitrary free parameter in the solution. The quantities p_1, p_2, and p_3 are then more simply expressed in terms of the single parameter p via

$$p_1 = \frac{-p}{p^2+p+1}\,, \quad p_2 = \frac{p(p+1)}{p^2+p+1}\,, \quad p_3 = \frac{p+1}{p^2+p+1}\,. \quad (19.3.8)$$

The condition $0 \leq p \leq 1$ on p then yields the conditions $-\frac{1}{3} \leq p_1 \leq 0$, $0 \leq p_2 \leq \frac{2}{3}$, and $\frac{2}{3} \leq p_3 \leq 1$. In the Kasner solution, the reference system expands in two directions (along the x_2- and x_3-axes) and contracts in the third (along the x_1-axis). The single exception is the degenerate case in which $p_1 = p_2 = 0$ and $p_3 = 1$. This, however, represents flat, four-dimensional spacetime satisfying $R_{iklm} = 0$. By a coordinate transformation, the metric (19.3.1) can in this case be brought to the Milne form. In the case of a general Kasner solution, of course, $R_{iklm} \neq 0$ even though $R_{ik} = 0$. Notice that the "vacuum," Kasner solutions do not include the isotropic solution $p_1 = p_2 = p_3$. Finally, a volume element changes as $V \sim t$ in the Kasner solution.

One can now find the domain of applicability of the solution obtained. Consider how the terms on the right-hand sides of equations (19.3.2–5) change upon time translations to the singularity, $t \to 0$, and to infinity, $t \to \infty$. On the left-hand sides of these equations enter quantities of the form $a^{-1}d^2a/dt^2$, $(a^{-1}da/dt)^2$, and so forth; each is of the order t^{-2}. Now suppose that the stress-energy tensor is hydrodynamic. The equation of state is $p = \beta\epsilon$, and the matter is at rest relative to the reference system. Then the quantities on the right-hand sides of the equations, $T_i{}^k$, are proportional to $V^{-(1+\beta)} \sim t^{-(1+\beta)}$. We see that for $0 \leq \beta < 1$, the exponent of t on the right-hand sides of equations (19.3.2–5) is less than 2, which is less than the exponent of t on the left-hand sides of the equations. Hence, upon approaching the singularity one can always neglect the terms on the right-hand sides by comparison with those on the left. Asymptotically,

as Lifshitz and Khalatnikov (1963*a*, *b*) emphasize, the solution does not depend on the presence of matter. One can therefore call this stage the "vacuum" stage.

Upon approaching $t \to \infty$, the matter terms decrease more slowly than t^{-2}. A moment of time is thus reached after which one cannot neglect these terms. The period of the "vacuum" solution is then over. We shall examine this late period of expansion in the next chapter.

19.4 A Comparison of the Newtonian and Relativistic Problems

For the very simple case when $p = 0$, the exact solution of equations (19.3.2–5) was obtained by Schücking and Heckmann (1958). We shall consider this solution first of all with the goal of comparing it with the solution of the analogous problem in Newtonian theory (see § 19.1). In addition, as we shall see later, this solution exhibits several basic properties of the solutions of more complicated problems. The Schücking-Heckmann solution is written in the form

$$\begin{aligned} a &= a_0 t^{p_1}(t + t_0)^{2/3-p_1}, \\ b &= b_0 t^{p_2}(t + t_0)^{2/3-p_2}, \\ c &= c_0 t^{p_3}(t + t_0)^{2/3-p_3}; \end{aligned} \tag{19.4.1}$$

$$\rho = \frac{1}{6\pi G}\frac{1}{t(t + t_0)}. \tag{19.4.2}$$

The conditions (19.3.7) and (19.3.8) on p_1, p_2, and p_3 also apply here. The expansion begins at $t = 0$ when $a = \infty$, $b = c = 0$, and $\rho = \infty$. In this solution, the vacuum stage occurs for $t \ll t_0$. For $t \gg t_0$, the solution is isotropized and passes over to a Friedmann solution.

We shall now compare the relativistic solution with the solution of the Newtonian problem presented above in § 19.1. To do this, we shall be examining, as was done there too, not expansion but contraction; we shall reverse the flow of time. Note that for this case we must replace t by $|t|$ and t_0 by $|t_0|$ in equations (19.4.1) and (19.4.2). As in the Newtonian solution, two qualitatively different cases are possible in the relativistic solution. The first case is related to a special, unique choice of the p_i: $p_1 = 0$, $p_2 = 0$, and $p_3 = 1$. The dependence of $a = b$ and of c on time is then completely analogous (in the limit $t \to 0$) to the case of a flattened ellipsoid of rotation in the Newtonian problem. If $p_i \neq 0$, however, the behavior of the solution near the singularity acquires entirely different character (see Fig. 55). For large $|t|$ (remember the time reversal!) we then have $d^2a/dt^2 < 0$, and $d^2c/dt^2 < 0$, analogous to the situation in the Newtonian problem. But for some value of t the sign of the relative acceleration along the x_1-axis changes, resulting in $d^2a/dt^2 > 0$. At the moment $t = -3p_1t_0/2$, the sign of the relative expansion rate along the x_1-axis also changes—

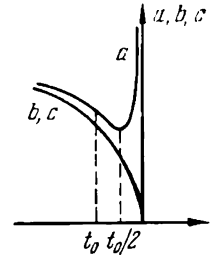


FIG. 55.—The collapse of homogeneous, anisotropic matter in general relativity theory. The graph displays the time evolution of the scale factors a, b, and c for the three principal axes, in the case $p_1 = -\frac{1}{3}$.

contraction changes into expansion—and a vacuum stage ensues. In the limit as $t \to \infty$, the distance between neighboring particles along the x_1-axis grows without bound.

The general solution has the features of the $p_1 \neq 0$ solution—not those of the $p_1 = 0$ solution. In the relativistic problem studied here, consequently, the contraction to a line is a general case, while contraction to a pancake is an exceptional case. Thus, the relativistic solution with a vacuum stage is qualitatively different from the solution of the Newtonian problem for an ellipsoid. The local properties of the relativistic solution are not obtained by means of a Newtonian examination of the problem, in contrast with the homogeneous, isotropic case.

The expansion along one axis and contraction along the two others during the vacuum stage is analogous to the tidal effect in the external field of a gravitating body. For example, the envelope of water surrounding the Earth (the oceans) also experiences relative accelerations in the gravitational field of the Moon. It is stretched out along the Earth-Moon line and contracted in the directions perpendicular to this line.

We shall now turn to the question of the reasons behind the lack of correspondence between the Newtonian and relativistic cosmological solutions. In Newtonian mechanics, one can change all linear scales if one also preserves all similarity relations. By so doing, one does not change the local properties of a solution, i.e., one does not change the time dependence of the quantities $\rho(t)$, $H_{ik}(t)$, and $\phi_{ik}(t)$. But why are these features not in accord with the properties of the relativistic cosmological solutions?

In the relativistic theory, the speed of light plays an important role. An exact relativistic solution of the problem of the motion of a finite volume filled with matter depends not only on ρ, H_{ik}, and ϕ_{ik} but also on dimensionless factors of the form $G\rho r^2/c^2$ and $H_{ik}r/c$, where r is a scale characteristic of the system.

The Newtonian solutions are a limiting form of exact relativistic solutions in the limit that the dimensionless factors mentioned above go to zero. However, the relativistic solutions must be realized in the opposite

limiting case when these factors tend to infinity. It is not surprising then that these solutions do not coincide with the Newtonian solutions.

A peculiarity of the problem is that in the special case where there is symmetry (corresponding to an isotropic solution), the relativity criteria come from formulae relating local quantities. This fact does not generalize to ellipsoidal solutions, however. The difference between the spherical and ellipsoidal cases is particularly evident in an examination of gravitational radiation.

In GTR, given motion of a spherical layer of matter, the field outside does not change and the field inside is zero and also does not change. Upon the motion of an ellipsoidal layer, however, the quadrupole moment of the mass distribution changes. Consequently, the radiation of a gravitational wave must take place.

Apparently, the field inside a moving ellipsoidal shell (equal to zero in the Newtonian approximation) also is nonzero in GTR. There occurs a sort of penetration of the gravitational wave into the cavity even though, strictly speaking, one cannot speak about a wave at distances less than ct, where t is a characteristic time of the motion. One speaks more appropriately of a variable gravitational field under such circumstances.

Consider a body for which the relativity criteria are not satisfied in the initial state—for which the factors mentioned earlier are small. If this body contracts, the density and the velocity grow. During succeeding stages, therefore, the motion in the finite-mass body can acquire the behavior characteristic of an anisotropic cosmological solution. Meanwhile, the variable gravitational field resulting from the motion of the external outer layers, a field having "tidal character" (stretching out matter in one direction and contracting it in the two other directions), determines the dynamics of the motion of the inner layers.

Consider finally the first, degenerate case with $p_1 = 0$, for which the relativistic solution is analogous to the Newtonian solution, and the second, general case with $p_1 \neq 0$, for which these solutions are not similar. Qualitatively, the distinction between them is that in the first case the gravitational field and the contraction rate of the volume V remain finite near the $\rho = \infty$ singularity in the Newtonian problem, while in the second case they grow without bound.

20 MATTER IN AN ANISOTROPIC COSMOLOGICAL MODEL

20.1 Isotropization of a Model with a Stress-Energy Tensor of the Pascal Form

Consider now the comparatively late stages of expansion of an anisotropic model when one can no longer neglect the terms on the right-hand side of the gravitational equations; these terms describe the gravitational effects of matter. The "vacuum" solution is not a valid description of these stages.

We shall first of all consider the matter to be at rest with respect to the reference system of the model. From the considerations mentioned in § 19.3, it is easy to determine the time when the vacuum solution ceases to be valid, given any hydrodynamic equation of state $p = \beta\epsilon$ for the material. This is the time for which the terms on the right-hand side of Einstein's equations become comparable to the terms on the left-hand side. The terms on the left-hand side are of the order of $a^{-1}d^2a/dt^2 \approx t^{-2}$, while the terms on the right side are of the order of $4\pi G\rho \approx 4\pi GAV^{-(1+\beta)} = 4\pi GAt^{-(1+\beta)}$. Here A is an arbitrary constant—a parameter of the problem—describing the amount of material in the model, and β is the coefficient in the expression $p = \beta\epsilon$. By equating these terms, we find the moment t_1 when the vacuum solution ceases to be valid. The result is

$$t_1 = (GA)^{1/(1-\beta)} . \tag{20.1.1}$$

The assertion that all terms on the left-hand side of the equations are of an order not less than t^{-2} is of course valid only if the exponent p_1 (and hence p_2 also) is not particularly small. In the opposite case, the answer depends also on p_1. When $|p_1| \ll 1$, the principal terms on the left-hand side of equations (19.3.2) and (19.3.3) are of the order of $|p_1|t^{-2}$, so we obtain

$$t_2 \approx (|p_1|/GA)^{-1/(1-\beta)} \tag{20.1.2}$$

for the time when the terms on the right-hand sides of equations (19.3.2–5) begin to influence the behavior. This formula generalizes formula (20.1.1) to the case of small $|p_1| \ll 1$.

As one can see, the solution rapidly approaches a Friedmann isotropic solution in the case where $|p_1|$ is not small. If $|p_1| \ll 1$, however, then with the replacement of contraction along the x_1-axis by expansion, following the

end of the vacuum stage, there ensues a prolonged stage with slow expansion along two axes and rapid expansion along the third axis. The solution is still highly anisotropic. By the time t_1 (given by formula [20.1.1]), however, the rate of expansion along the three axes is equalized, and isotropization ensues. Thus for any p_1, t_1 is always the time of "isotropization" of an anisotropic solution.

The physical reason for the comparatively weak gravitational effect of the material on the dynamics of the expansion in the case $|p_1| \ll 1$ is discussed in the last paragraph of the previous section.

With respect to applications, the case of the hot model, when the equation of state is $p = \epsilon/3$, is evidently of most interest. Exact solutions for this problem were obtained by Kompaneets and Chernov (1964), Doroshkevich (1965), and Ruban (private communication). We shall not present the exact form of the solution here but shall point out a convenient approximate form. This form approximates the solution well over the whole interval of time $0 \leq t < \infty$, and gives the exact asymptotic expressions as $t \to 0$ and as $t \to \infty$. It can be written in the following fashion:

$$a_i = a_{0i} t^{p_i} (t + t_*)^{1/2 - p_i} ; \tag{20.1.3}$$

$$\rho = \frac{3}{32\pi G} t^{-4/3} (t + t_*)^{-2/3} . \tag{20.1.4}$$

Here a_i are the scale factors for the three axes, and the p_i satisfy the relations (19.3.7). For $t \to 0$, we obtain the Kasner solution, as we must. For $t \to \infty$, the solution passes over to a Friedmann solution. The isotropization time is $t_1 \approx t_*$, and it is convenient to take this time as an arbitrary parameter of the problem.

20.2 The Effect of a Spatially Anisotropic Stress-Energy Tensor on a Cosmological Model

Up to now we have only considered cosmological models with a hydrodynamical stress-energy tensor $(T_{ik})_{\text{hydro}} = (\epsilon + p) u_i u_k + p g_{ik}$. This tensor is spatially isotropic; the Pascal law is fulfilled for the pressure. As is well known, however, this condition is by no means always fulfilled. For example, if a homogeneous magnetic field is oriented along the z-axis, the stress-energy tensor can be written in the form

$$-T_1^{\,1} = -T_2^{\,2} = T_3^{\,3} = \frac{H^2}{8\pi} , \qquad T_0^{\,0} = \frac{H^2}{8\pi} , \tag{20.2.1}$$

where H is the magnetic field strength. This tensor is anisotropic. There is a tension or "negative pressure" along the z-axis but a positive pressure along the x- and y-axes.

A directed flux of collisionless relativistic particles (photons, for example) moving along the x-axis equally in both directions, say, can also serve as an example. For such a flux, the stress-energy tensor takes the form

$$T_1^1 = -\epsilon\,, \quad T_0^0 = \epsilon\,. \tag{20.2.2}$$

The quantity ϵ is the energy density of the particles, and all of the other components of T_i^k vanish. Here the pressure is positive, and equal to the energy density, but it only acts in the direction of the x-axis. Along the two other axes, the pressure is zero.

We emphasize that in both cases mentioned, there is no directed energy flow relative to the frame of reference; $T_0^i = 0$ for all i.

We shall see later that upon an examination of problems involving anisotropic cosmological models, the emergence of spatial anisotropy in the stress-energy tensor is inevitable. In particular, the examination models (20.2.1) and (20.2.2) studied above are not artificial. Tensors of such a form are actually encountered. In this section, we shall not be concerned with cases of directed energy flows, when $T_k^i \neq 0$, even though these cases are important for cosmology. Indeed, the cases with $T_k^i \neq 0$ require a special approach; they will be considered further in § 20.5. In this section the nature of the effect of spatial anisotropy on the expansion and on the acceleration of matter in various directions is examined.

We emphasize first of all that, as a consequence of homogeneity, there is no pressure gradient at all nor any force that would be connected with such a gradient. The effect of the pressure anisotropy can only be gravitational. As is well known, the gravitational field in GTR can also depend on pressure (see § 2.1 and, for more detail, § 1.5 and § 1.6 in Volume 1). An isotropic, positive pressure increases the deceleration of the expansion. An anisotropic pressure must give rise to anisotropy in the accelerations of the matter deformation. Indeed it is apparent from the form of equations (19.3.2-5) just how the anisotropy of T_i^k influences the relative acceleration. We shall really consider two solutions which we shall call 1 and 2.

For both solutions, suppose that at some instant of time all first derivatives along corresponding axes are equal: $(a^{-1}da/dt)_1 = (a^{-1}da/dt)_2$; $(b^{-1}db/dt)_1 = (b^{-1}db/dt)_2$; and $(c^{-1}dc/dt)_1 = (c^{-1}dc/dt)_2$. Suppose further that we have $(T_0^0)_1 = (T_0^0)_2$, $(T)_1 = (T)_2$ as conditions on the energy density and the trace of the stress-energy tensor. Finally, assume that the tensor T_α^β is isotropic, $T_1^1 = T_2^2 = T_3^3$, for the first solution and anisotropic for the second. How does this affect the acceleration? We find from equations (19.3.2-5) that

$$\begin{aligned}
(a^{-1}d^2a/dt^2)_2 - (a^{-1}d^2a/dt^2)_1 &= -8\pi G[(T_1^1)_2 - (T_1^1)_1]\,,\\
(b^{-1}d^2b/dt^2)_2 - (b^{-1}d^2b/dt^2)_1 &= -8\pi G[(T_2^2)_2 - (T_2^2)_1]\,,\\
(c^{-1}d^2c/dt^2)_2 - (c^{-1}d^2c/dt^2)_1 &= -8\pi G[(T_3^3)_2 - (T_3^3)_1]\,.
\end{aligned} \tag{20.2.3}$$

Thus, the anisotropy emerging in the relative acceleration is proportional to the anisotropy in T_α^β and is opposite in sign. It is evident that the tensor T_{ik} can have a noticeable effect on the expansion only when there are on the right-hand sides of equations (19.3.2–5) terms comparable in order of magnitude to the terms on the left-hand side. Indeed, this is also true for an isotropic T_{ik} in anisotropic models. Generally speaking, however, an anisotropic tensor T_{ik} cannot guarantee eventual isotropization of the solution in the absence of other factors which also lead to isotropization (see more about this below).

To illustrate these remarks, we shall study the anisotropic cosmological solutions with the stress-energy tensors (20.2.1) and (20.2.2) presented earlier.

The first solution, with a homogeneous magnetic field and the metric (19.3.1), has been considered, for example, by Rosen (1964). Suppose the field is directed along the z-axis and that $W = H^2/8\pi$. The solution can then be written in parametric form:

$$a = a_0\xi^p(\xi + \xi^{-1}), \quad b = b_0\xi^{1/p}(\xi + \xi^{-1}), \quad c = c_0(\xi + \xi^{-1})^{-1};$$
$$dt = \xi_0\xi^{(p^2-p+1)/p}(\xi + \xi^{-1})d\xi; \quad \xi_0 = \frac{(\kappa W)^{1/2}}{2}\frac{ab}{a_0b_0} = \text{const.} \tag{20.2.4}$$

Here, $0 < p \leq 1$, $0 < \xi < \infty$, and κ is the Einstein gravitational constant. Note that $W^{1/2} \sim H \sim (ab)^{-1}$ since H is proportional to the area.

Near the singularity, $\xi \to 0$, the solution has the asymptotic form

$$a = a'_0t^{p_1}, \quad b = b'_0t^{p_2}, \quad c = c'_0t^{p_3}, \tag{20.2.5}$$

where p_1, p_2, and p_3 satisfy relations analogous to the relations (19.3.8), namely,

$$p_1 = \frac{p(p-1)}{p^2-p+1}, \quad p_2 = \frac{1-p}{p^2-p+1},$$
$$p_3 = \frac{p}{p^2-p+1}; \quad 0 < p \leq 1. \tag{20.2.6}$$

Evidently p_1, p_2, and p_3 satisfy relation (19.3.7) as a consequence, so the asymptotic solution (20.2.5) is of the "vacuum" form. The exponent p_1 is never positive, $0 \leq -p_1 \leq \frac{1}{3}$, while p_2 and p_3 are never negative, with both $p_2 > p_3$ and $p_3 > p_2$ possible. Since the solution is of a "vacuum" nature for $t \to 0$, the magnetic field does not affect the solution in this limit. Actually, the field energy grows as

$$H^2 \sim t^{-2(p_1+p_2)} \tag{20.2.7}$$

in this case. But as a consequence of relations (20.2.6), we have $2(p_1 + p_2) < 2$. One can therefore neglect the terms on the right-hand side of the gravita-

tional equations by comparison with the terms on the left-hand side in this limit. The latter are proportional to t^{-2} here.

We emphasize that the problem with a magnetic field does not admit a solution with a negative p_3 (the field is along the z-axis, which p_3 characterizes). For such a solution, the field energy would grow more rapidly than t^{-2} as $t \to 0$ and one would then not be able to neglect the magnetic field terms near the singularity.

When, in the course of the expansion, the vacuum stage ends, the right-hand sides of the gravitational equations begin to influence the solution. But this influence in the solution under consideration differs from the influence that would obtain given an isotropic tensor $T_{\alpha\beta}$. Let us now study the nature of this influence. Compare the effect of the tensor $T_{\alpha\beta}$ describing a magnetic field with the effect of this tensor when the material is such that the tensor is isotropic. Assume moreover in both cases the same energy density ϵ and the same average pressure $\langle p \rangle = \epsilon/3$ (the same value of $T_1{}^1 + T_2{}^2 + T_3{}^3$); hence $T = T_i{}^i$ is the same for both. The tensor $(T_{\alpha\beta})_{\text{isot}}$ evidently describes a relativistic gas. Substituting for $T_{\alpha\beta}$ the magnetic field form (20.2.1), and the relativistic gas form into equations (20.2.3), we get

$$\Delta(a^{-1}d^2a/dt^2) \equiv (a^{-1}d^2a/dt^2)_2 - (a^{-1}d^2a/dt^2)_1 = \Delta(b^{-1}d^2b/dt^2) = 8\pi G(\tfrac{2}{3})\epsilon\,, \quad \Delta(c^{-1}d^2c/dt^2) = -8\pi G(\tfrac{4}{3})\epsilon\,. \tag{20.2.8}$$

Most pronounced of all is the effect of the anisotropy of $T_{\alpha\beta}$ on $c^{-1}d^2c/dt^2$, an effect which increases the modulus of the negative quantity $c^{-1}d^2c/dt^2$. As a result, the slowed expansion along the z-axis goes over into contraction. Along the x-axis, the contraction changes into expansion, just as in the isotropic case. As a whole, the solution again assumes a "vacuum" form as $t \to \infty$. It again takes the form (20.2.5), but p_1 and p_3 must then be exchanged in relations (20.2.6).

Of course, the problem studied is a model and for $t \to \infty$; the solution is in any event not applicable to a description of the real Universe. As we shall see in § 20.5, however, for a finite time interval the solution is possibly relevant to an analysis of processes taking place at the beginning of the expansion.

Now let us present the solution for relativistic particles moving along the z-axis, for which $T_{\alpha\beta}$ satisfies

$$-T_0{}^0 = T_3{}^3 = -\epsilon\,, \quad T_1{}^1 = T_2{}^2 = 0\,. \tag{20.2.9a}$$

We have

$$a = a_0\xi^{1/2-p}\,, \quad b = b_0\xi^{1/2+p}\,, \quad c = c_0\xi^{p^2-1/4}e^{\xi}\,; \quad dt = t_0\xi^{p^2-1/4}e^{\xi}d\xi\,; \quad \kappa\epsilon = t_0^{-2}\xi^{-1/2-2p^2}e^{-2\xi}\,. \tag{20.2.9b}$$

Here a_0, b_0, c_0, and t_0 are constants, $0 \le p \le \frac{1}{2}$; and κ is the Einstein gravitational constant.

For $t > 0$, the solution is asymptotically of the "vacuum" form with a negative exponent p_3 characterizing contraction along the z-axis. For $t \to \infty$, the solution again becomes of the "vacuum" form, but with the expansion along the z-axis following the law $c \sim t$ and with a logarithmic slowing along the x- and y-axes. The reader can easily carry out a qualitative analysis of the effect of anisotropy in $T_{\alpha\beta}$ on the solution in this case by himself by analogy with the preceding analysis.

20.3 Cosmological Models with a Homogeneous Magnetic Field

Until recently, the problem of the spontaneous emergence of a magnetic field in galaxies encountered serious difficulties (see Part III). Recently, much research has appeared, both theoretical and studies based upon an interpretation of observational data, in which a primordial magnetic field that exists prior to the appearance of galaxies is studied.

At the present time, the theory of the generation and amplification of a magnetic field probably permits one to explain the emergence of magnetic fields in galaxies. The primary reason for introducing a primordial magnetic field into theoretical cosmology has thus vanished. Nevertheless, of course, there has been no demonstration whatsoever of the absence of a primordial intergalactic field whose strength is $H \sim 10^{-7}$ gauss or less today. Moreover, articles have appeared (see, for example, Reinhardt and Thiel 1970; Reinhardt and Roberts 1972; Reinhardt 1972) based on an interpretation of the observational data in which views are expressed in favor of the existence of such a field.

In this section, therefore, we shall consider cosmological models with a homogeneous magnetic field. In passing, we mention here several authors who have worked on this problem (this list is of course incomplete): Hoyle (1958), Brachmachary (1965), Rosen (1964), Zel'dovich (1965*b*, 1969), Khalatnikov (1965, 1967), Doroshkevich (1965), Thorne (1967), Misner (1968), Jacobs (1969) and Vajk (1969).

The most important property of such models consists of the existence of a magnetic field in spite of the complete absence of an electric current. That this is possible follows formally from the homogeneity of the field: $\nabla \times \mathbf{H} = 0$. Recall that a finite body possesses a surface current if this condition is imposed upon the field inside. Now the GTR equations are compatible with the Maxwell equations, so the conditions corresponding to a geometrically "frozen-in" field can be satisfied independent of the conductivity of the matter—or even in the absence of matter. The condition is that the flux passing through a contour consisting of given particles of the matter be conserved.

A time-dependent magnetic field is accompanied by an electric field. If, however, the magnetic field is frozen in, an observer moving with the matter

is incapable of observing an electric field in his system. Indeed, this is also a consequence of the frozen-in character.

We shall introduce this property into GTR for the metric (19.3.1), where we now use the notation $x = x_1$, $y = x_2$, and $z = x_3$. The Maxwell equations are written in the form

$$\frac{1}{\sqrt{-g}} \frac{\partial}{\partial x_i} [(\sqrt{-g}) F^{ik}] = 0 \qquad (20.3.1)$$

since the current vanishes by virtue of the homogeneity of the field. Now, assume that the homogeneous magnetic field is oriented along the z-axis. Then the particular solution for the aforementioned Maxwell equations becomes

$$H_x = H_y = 0\,, \quad H_z = \frac{\text{const.}}{ab}\,; \quad E_x = E_y = E_z = 0\,. \quad (20.3.2)$$

The magnetic field does not give rise to an electric field in the comoving space because the lines of force of the field do not move relative to the reference system.

Another solution of Maxwell's equations, one independent of the previous solution, involves a frozen-in electric field directed along any of the axes. For example, the solution for a field along the z-axis is

$$E_x = E_y = 0\,, \quad E_z = \frac{\text{const.}}{ab}\,; \quad H_x = H_y = H_z = 0\,. \quad (20.3.3)$$

It is evident that this solution describes a frozen-in electric field.

Formally, all of the models and conclusions about them discussed in this section remain valid if the homogeneous magnetic field is replaced by an electric field. However, a homogeneous intergalactic electric field must give rise to an electric current of charged particles and then rapidly decay. Indeed, there is at the present time no basis whatsoever to presume the presence of an overall electric field (in distinction to the assumption of an overall magnetic field). A model with an overall electric field therefore bears hardly any relation at all to reality.

Let us now return to the model with a magnetic field and expanding matter. Charged particles moving relative to the matter are deflected by the field, but their energy, as measured by an observer at the particles' instantaneous location, can only decrease in the course of the expansion. We previously found this result in the study of other cases.

The equations for the general problem of a magnetic model of the Universe are obtained from equations (19.2.4) by adding to the right-hand side, where the magnetic field tensor appears, the stress-energy tensor for ordinary material. For a relativistic gas this addition involves the quantities $(T_0{}^0)_{\text{mat}} = \epsilon$ and $(T_1{}^1)_{\text{mat}} = (T_2{}^2)_{\text{mat}} = (T_3{}^3)_{\text{mat}} = -p = -\epsilon/3$.

Referring for details of the calculations to the work done by Doroshkevich (1965) and by Zel'dovich (1969), we shall present here qualitative

considerations and conclusions. Since we already know about separate anisotropic solutions with ordinary matter only (see § 20.1) and with a magnetic field only (see § 20.2), it will not be difficult to obtain the combined solution.

First of all, as we saw in § 20.2, a magnetic field does not permit a solution in which, as $t \to 0$, there is a negative exponent in the time dependence of the scale factor along the field direction. This property is preserved in the combined solution under consideration. The physical meaning of this property is the same as it was in § 20.2: namely, if both scale factors in the directions orthogonal to that of the field were to decrease for $t \to 0$, then the energy density of the field, $\epsilon \sim H^2$, would tend to infinity faster than t^{-2}. The gravity caused by the magnetic field would then change the nature of the solution, converting it to a vacuum solution as $t \to 0$, with expansion along one of the axes orthogonal to the field direction.

The end of the vacuum stage is signaled when one of the terms on the right-hand side of equation (20.2.4), $(T_i^k)_{\text{field}}$ or $(T_i^k)_{\text{mat}}$, becomes comparable with t^{-2}. If the first term to become comparable with t^{-2} is the $(T_i^k)_{\text{mat}}$ term, the one related to the ordinary matter, then this signifies that the energy density of the magnetic field is less than the energy density of the ordinary matter up to the end of the vacuum stage. The field then turns out to have no influence at all on the dynamics of expansion during any stage, so the solution coincides with that described in § 20.1.

Finally, if the first term to become comparable with t^{-2} is the term $(T_i^k)_{\text{field}}$, the one related to the magnetic field, then the field influences the expansion substantially. It transforms the solution, characterized by one type of vacuum asymptotic behavior, to a solution with another type of such behavior (see § 20.2). Only after the gravity due to the material becomes important does the solution isotropize and approach a Friedmann solution.

We shall now examine in more detail the process of isotropization of a solution (Zel'dovich 1969). Namely, we shall examine the stage during which the solution is almost isotropic and the influence of the field anisotropy is small. In order to obtain quantitative estimates, we shall consider an axisymmetric solution with the metric

$$ds^2 = dt^2 - a^2(t)(dx^2 + dy^2) - b^2(t)dz^2 \tag{20.3.4}$$

and with the field oriented along the z-axis. We use the notation $\alpha = a^{-1}da/dt$ and $\beta = b^{-1}db/dt$. The Einstein equations then take the form (with $8\pi G = 1$ and $c = 1$)

$$(a^2b)^{-1}\frac{d}{dt}(a^2b\alpha) = \frac{\epsilon}{3} + W, \quad (a^2b)^{-1}\frac{d}{dt}(a^2b\beta) = \frac{\epsilon}{3} - W \cdot$$
$$W = \frac{H^2}{8\pi} = \text{const.}\, a^{-4}, \quad \epsilon = \text{const.}\,(a^2b)^{-4/3}. \tag{20.3.5}$$

If one eliminates a and b, these equations become

$$\frac{d\alpha}{dt} + \alpha(2\alpha + \beta) = \frac{\epsilon}{3} + W\,, \quad \frac{d\beta}{dt} + \beta(2\alpha + \beta) = \frac{\epsilon}{3} - W\,;$$
$$\frac{d\epsilon}{dt} = -\tfrac{4}{3}(2\alpha + \beta)\epsilon\,, \quad \frac{dW}{dt} = -4\alpha W\,. \tag{20.3.6}$$

Using this notation, an isotropic solution is

$$\alpha = \beta = (2t)^{-1}\,; \quad W = 0\,, \quad \epsilon = \frac{3}{4t^2}\,. \tag{20.3.7}$$

Let us distinguish the quantity $n = (2\alpha + \beta)/3$ characterizing isotropic expansion and, on the other hand, the quantities $r = (\alpha - \beta)/n$ and $q = W/\epsilon$ characterizing anisotropy. Employing these quantities, we obtain the equations

$$3\frac{dn}{dt} + 9n^2 = \epsilon(1 + q)\,, \quad r\frac{dn}{dt} + n\frac{dr}{dt} + 3n^2 r = 2q\epsilon\,;$$
$$\frac{d\epsilon}{dt} = -4n\epsilon\,, \quad \frac{dq}{dt} = -\tfrac{4}{3}rnq\,. \tag{20.3.8}$$

In addition, one more relation holds, namely, the integral of the equations written above. It is

$$(2\alpha + \beta)^2 - 2\alpha^2 - \beta^2 = 2\epsilon + 2W\,,$$

or, in terms of the r, n, and q notation,

$$3n^2\left(1 - \frac{r^2}{9}\right) = \epsilon(1 + q)\,. \tag{20.3.9}$$

The relation (20.3.9), the integral of the system (20.3.8), permits one to eliminate one of the differential equations, the first for example.

The system of equations written above in principle allows one to construct a cosmological solution with any magnetic field specified in advance with any expansion anisotropy at a given instant, and with a given energy density ϵ. Having specified ϵ, q, r, and t_*, we find the corresponding n from the last formula. We then integrate the equations into the past and into the future. The only limitation is that the RD period is considered (when $p = \epsilon/3$, i.e., for $z \gtrsim 2 \times 10^4$ or $t > t_2 \approx 10^{12}$ s).

Now, from the problem of a flat model with anisotropic expansion and a magnetic field, we pass to the specific case of a weak field and small anisotropy, $q < 1$ and $r < 1$. We require that these conditions be fulfilled for the duration of an interval from some t_1 to t_2. One can choose the beginning of the interval t_1 as either the Planck time, so $t_1 = 10^{-43}$ s, or the time of the freeze-out in the equilibrium between neutrons and protons ($t_1 = 1$ s). The

choice depends on the formulation of the problem. Do we require that the model be as close as is possible to a Friedmann solution—as ordered as is possible over the entire evolution? Or are we satisfied with an absence of influence by the magnetic field vis-à-vis predictions of the theory describing the relative chemical composition of primordial matter?

In any event, the mathematics are significantly simplified, given the restriction to $q < 1$ and $r < 1$. The energy density ϵ and the overall expansion rate n which follow from these restrictions differ little from the isotropic values, $n = (2t)^{-1}$ and $\epsilon = 3/(4t^2)$. We now substitute these quantities into the equations for small r and small q and obtain

$$\frac{dr}{dt} = -\frac{r}{2t} + \frac{3q}{t}; \quad \frac{dq}{dt} = -\frac{2}{3t} rq . \tag{20.3.10}$$

In the absence of a magnetic field ($q = 0$), the expansion anisotropy rapidly decreases: $r = r_0(t_0/t)^{1/2}$. On the other hand, a magnetic field acts so as to "conserve" the expansion anisotropy, just as do fluxes of neutrinos and gravitons.

It is easy to find a quasi-stationary solution. Given the presence of a magnetic field, r evidently tends not to zero but to that value for which $dr/dt = 0$, i.e., $r \to 6q$. We substitute this value of r into the second equation and find

$$\frac{dq}{dt} = -\frac{4}{t} q^2; \quad q(t) = \frac{q_1}{1 + 4q_1 \ln (t/t_1)} . \tag{20.3.11}$$

Consequently, we obtain the inequality

$$q(t_2) < [4 \ln (t_2/t_1)]^{-1} . \tag{20.3.12}$$

For $t_1 = 10^{-43}$ s and $t_2 = 10^{12}$ s, $q = q' < 2 \times 10^{-3}$. For $t_1 = 1$ s and $t_2 = 10^{12}$ s, $q = q'' < 10^{-2}$. The ratio of the magnetic field energy to the RR energy remains almost constant in the course of the further evolution—right up to the present. The present energy density ϵ_r is 4×10^{-13} ergs cm^{-3}. In the two variants, the one with $q' < 2 \times 10^{-3}$ and that with $q'' < 10^{-2}$, we obtain $H = H' = 1.5 \times 10^{-7}$ gauss and $H = H'' = 3 \times 10^{-7}$ gauss, respectively, for the magnetic field.

The conjectured primordial (cosmological) magnetic field is substantially less—of the order of 10^{-10} to 10^{-9} gauss. Such an estimate follows from the magnitude of the magnetic field of galaxies, which is of the order of 10^{-6} gauss. But the density of the matter contained in galaxies is 10^5–10^6 times greater than the average density of matter in the Universe. In the very simple case of isotropic contraction, the field grows in proportion to $\rho^{2/3}$. The field of galaxies is estimated to be 10^3–10^4 times greater than the primordial field even if no account is taken of the strengthening of the field due to differential rotation and the dynamo mechanism (see § 15.7).

The conclusion is thus that a primordial cosmological magnetic field is necessary for an explanation of the magnetic field of galaxies. It is "inserted" in the isotropic Friedmann model and otherwise causes only small changes in the properties of that model—even over the whole duration of evolution.

It is also easy to convince oneself that such a magnetic field, of the order of 10^{-10} to 10^{-9} gauss, gives rise to a fairly small expansion anisotropy, one corresponding to $q = 5 \times 10^{-8}$ to $q = 5 \times 10^{-6}$. Consequently, the expansion anisotropy is in the range $r \sim 3 \times 10^{-7}$ to $r \sim 3 \times 10^{-5}$ during the RD stage. Later, when $\rho_m > \rho_r$, the value of r for a stationary solution decreases, since $r = 6\epsilon_H/(\epsilon_r + \rho_m c^2) = 3H^2/[4\pi(\epsilon_r + \rho_m c^2)]$ in order of magnitude.

The RR temperature anisotropy is conveniently expressed in terms of the parameter r. In fact, suppose that the model becomes transparent at the instant t_0, that the temperature is equal to T_0 at this instant, and that the radiation is isotropic. We find that rays propagating along the x- or y-axis are thereafter characterized by the temperature

$$T' = T_0 \frac{a_0}{a} = T_0 \exp\left(-\int \alpha dt\right) . \tag{20.3.13}$$

For rays propagating along the z-axis, we have

$$T'' = T_0 \frac{b_0}{b} = T_0 \exp\left(-\int \beta dt\right) . \tag{20.3.14}$$

Consequently,

$$\begin{aligned} \frac{\Delta T}{T} = \frac{T' - T''}{T'} &= \frac{\exp\left(-\int \alpha dt\right) - \exp\left(-\int \beta dt\right)}{\exp\left(-\int \alpha dt\right)} \\ &= 1 - \exp\left[+\int (\alpha - \beta) dt\right] \approx \int (\beta - \alpha) dt \\ &= -\int r n dt \approx -\int \frac{r}{2t}\, dt = -\frac{1}{2} \int r d(\ln t) . \end{aligned} \tag{20.3.15}$$

The RR anisotropy corresponding to a field strength today of 10^{-10} to 10^{-9} gauss turns out to satisfy $\Delta T/T < 10^{-6}$. This is far beyond the limits of accuracy of present observations and less than the possible anisotropy due to other causes.

The hypothesis of a primordial magnetic field therefore does not contradict observations, nor does it contradict the assumption of nearly Friedmannian behavior just after the singularity. A desire to do without a primordial field—the desire to explain the galactic fields by means of the dynamo mechanism—and a skeptical attitude toward this hypothesis are both connected with intuitive, almost aesthetic motives and with foresight regarding future singularity theory. It does not seem likely that a magnetic field which destroys the symmetry between left and right, would spontaneously appear in this theory. However, one must keep in mind the precariousness and subjectivity of such aesthetic criteria.

30.1 Perturbations in a Homogeneous, Anisotropic (HA) Universe

In the preceding part of the book, questions connected with fluctuations in the isotropic and homogeneous Friedmann model were considered. If, however, the initial stages of the cosmological expansion are not Friedmannian, then the laws describing the growth of fluctuations naturally are different. In this section we shall study the problem of the growth of density perturbations in anisotropically expanding matter that is on the average at rest relative to a synchronous reference system. We shall also study the change in amplitude of gravitational and acoustic waves (Lifshitz and Khalatnikov 1963*a*, *b*; Doroshkevich 1966; Doroshkevich, Zel'dovich, and Novikov 1971).

In the part of the book concerned with the gravitational instability of the Friedmann model, it was shown in detail that the cause of the growth of density perturbations leading to the Jeans instability is the self-gravitation of the perturbations. Denser blobs of matter produce stronger gravitational fields, which in turn enhance the growth of the blobs. The growth of perturbations in an anisotropic model is of an entirely different character.

The aim of this section is to show that the growth of density perturbations in the matter in an anisotropic, expanding Universe during the vacuum stage is substantial, but it is a kinematic effect. The kinematic effect results from the motion of the matter in the gravitational field that is specified by the solution of the empty-space Einstein equations. It is a further aim to find the laws describing the growth of the perturbations of the matter density. A clear understanding of the process of the growth of fluctuations permits one to move on to an examination of the final, large density inhomogeneity that occurs in several particular cases of the problem.

An examination of inhomogeneous perturbations (perturbations depending on the coordinates) in an HA Universe is of great interest. One can consider such an examination as a first approximation to the solution of the general problem of an inhomogeneous and anisotropic Universe. However, the nonlinearity of the GTR equations, the complexity of the physical processes, and mathematical difficulties make it impossible to solve the general problem straightforwardly at this time. One must instead acquire information about the character of the solution by studying special cases, among which exact solutions play an important role. As a rule these special cases are degenerate, possessing, for example, the type of invariance characteristic of spherical symmetry (invariance with respect to the rotation group) or of spatial homogeneity (invariance with respect to the translation group).

The large number of weakly perturbed exact solutions are of much greater utility (coinciding with the utility of a general solution) than the few exact solutions of special cases since the perturbations remove the

degeneracy and do not possess the invariance of the exact solutions. Moreover, as long as the perturbations are small, they satisfy linear equations and, after a Fourier decomposition with respect to space, lead to ordinary differential equations for the Fourier amplitudes. The small-perturbation problems therefore combine comprehensibility of the solutions and mathematical simplicity with the great generality of initial conditions.

Small density perturbations of the matter on the background of a flat, anisotropic model of the Universe are considered below. The physical interpretation of the results, which show that the perturbation growth is a kinematical effect, can be useful in an examination of more complicated cases. Such cases include, in particular, problems concerned with the evolution of perturbations during the stage of finite, but not small, density inhomogeneities and of finite, but not small, velocity perturbations.

Consider the HA model with ordinary matter. As has been repeatedly noted, the gravitation caused by the matter does not play a role during the early, vacuum stage of expansion of an anisotropic model.

Let us give a Newtonian description of the situation. Locally, an observer senses tidal forces, by comparison with which the gravitational interaction with the nearby volume elements is negligible. Additionally, a volume element of the comoving coordinate system contracts along one axis and expands along the two others. This is because the scale factors $a \sim t^{p_1}$, $b \sim t^{p_2}$, and $c \sim t^{p_3}$ for the three axes have exponents that satisfy (see eq. [19.3.8])

$$-\tfrac{1}{3} \leq p_1 \leq 0 \leq p_2 \leq \tfrac{2}{3} \leq p_3 \leq 1 \qquad (20.4.1)$$

in the general nondegenerate case. For later convenience, we shall use the notation $\alpha = -p_1$, with $0 \leq \alpha \leq \frac{1}{3}$.

Consider at first particles at rest with respect to the comoving coordinates, i.e., with the constant coordinates ξ, η, and ζ. Their "laboratory" coordinates are

$$x = a\xi\,, \qquad y = b\eta\,, \qquad z = c\zeta\,.$$

Take pairs of particles located along some axis, for example the two particles at $(\xi_1, 0, 0)$ and $(\xi_2, 0, 0)$ or the two at $(0, \eta_1, 0)$ and $(0, \eta_2, 0)$. The physical separation between the two particles is $x_{12} = a(\xi_1 - \xi_2)$ and $y_{12} = b(\eta_1 - \eta_2)$, respectively.

In a general situation, the relative acceleration of two particles is

$$\begin{aligned}
\frac{d^2x_{12}}{dt^2} &= p_1(p_1 - 1)t^{-2}x_{12} = \alpha(\alpha + 1)t^{-2}x_{12} = \left(a^{-1}\frac{d^2a}{dt^2}\right)x_{12}\,; \\
\frac{d^2y_{12}}{dt^2} &= -p_2(1 - p_2)t^{-2}y_{12} = \left(b^{-1}\frac{d^2b}{dt^2}\right)y_{12}\,, \qquad (20.4.2) \\
\frac{d^2z_{12}}{dt^2} &= -p_3(1 - p_3)t^{-2}z_{12} = \left(c^{-1}\frac{d^2c}{dt^2}\right)z_{12}\,.
\end{aligned}$$

In the Newtonian interpretation, such a relative acceleration results from the action of the gravitational potential, which satisfies the equations

$$\frac{d^2\phi}{dx^2} = -a^{-1}\frac{d^2a}{dt^2}, \quad \frac{d^2\phi}{dy^2} = -b^{-1}\frac{d^2b}{dt^2}, \quad \frac{d^2\phi}{dz^2} = -c^{-1}\frac{d^2c}{dt^2}. \quad (20.4.3)$$

In agreement with equation (19.3.5), we then obtain $\nabla^2\phi = 4\pi G(\epsilon + T_\alpha^\alpha)/c^2$ for the general case. For nonrelativistic matter, $T_\alpha^\alpha < \epsilon$ and $\epsilon/c^2 = \rho$, so $\nabla^2\phi = 4\pi G\rho$. During the vacuum stage, $\nabla^2\phi = 0$, with the separate terms $\partial^2\phi/\partial x^2$ for example, satisfying $\partial^2\phi/\partial x^2 \gg \nabla^2\phi$. Inasmuch as $\nabla^2\phi = 0$, which corresponds to the Newtonian equation for empty space, a local observer can say that the gravitational forces are not connected with the matter present and in this sense can call these forces tidal. From the viewpoint of relativistic theory, the gravitational field in this model is a free gravitational field like that of a gravitational wave with infinitely large wavelength. Such a field does not have its own matter source.

The tidal forces push apart particles originally at rest along the x-axis in the comoving reference system, thereby slowing the contraction, and draw comoving particles together along the y- and z-axes, thereby slowing the expansion. Particles moving at an arbitrary velocity experience the same gravitational forces as long as their velocity is nonrelativistic.

We shall now turn to a consideration of perturbations in the model at hand. We shall analyze only the Kasner, vacuum stage because the subsequent stage rapidly transforms into a Friedmann model, for which the behavior of the perturbations is known. As was already mentioned, the gravitational interaction of the matter does not play a role during the vacuum stage in the behavior of the perturbations just as it does not play a role in the behavior of unperturbed motion. Let us show this. We obtain an estimate of the role of the gravitational interaction of the matter by taking the instantaneous value of the density increment ($\delta = \Delta\rho/\rho$) according to the Jeans formula:

$$\frac{d(\ln\delta)}{dt} = (4\pi G\rho)^{1/2}, \quad \delta \sim \exp\left[\int (4\pi G\rho)^{1/2}dt\right].$$

In the case of dust, $\rho = \rho_m$, while in the case of radiation, $\rho = \rho_r$; we shall use the formulae $\rho_m = At^{-1}$ and $\rho_r = Bt^{-4/3}$ (recall that $\rho_m \sim V^{-1}$ and $\rho_r \sim V^{-4/3}$ and that $V \sim t$ in the Kasner solution). We shall express the constants A and B in terms of the t_1, the isotropization time (the time when the solution assumes a Friedmannian behavior); thus

$$At_1^{-1} = (6\pi Gt_1^2)^{-1}, \quad Bt_1^{-4/3} = 3(32\pi Gt_1^2)^{-1}.$$

We then find that during the time from the singularity, at $t = 0$, to t_1, $\ln\delta$ grows to the finite value

$$\int_0^{t_1} (4\pi G\rho)^{1/2}dt = \begin{cases} 2(2/3)^{1/2} & \text{for} \quad \rho = \rho_m, \\ (3/2)^{3/2} & \text{for} \quad \rho = \rho_r. \end{cases}$$

For an isotropic model, such an integral diverges at the lower limit. By comparing the contribution of the gravitational interaction with these results, one can infer that it is not substantial. In the absence of a gravitational interaction, it is impossible to talk about the gravitational instability of a homogeneous world.

However, this of course does not mean that all perturbations either decay or fail to grow. During the vacuum period, there are growing perturbations, with the law describing this growth even more pronounced than the perturbation growth that occurs due to the Jeans instability in the isotropic Friedmann model.

The growth of perturbations in an anisotropic model has a substantially non-Jeans-like nature. This growth is related to the contraction along one axis (the x-axis here). We consider the "dust" case ($p = 0$). In the frame of reference (x, y, z) the "dust" has a peculiar velocity v. Since gravitational forces are not important, the perturbations grow for kinematic reasons. In this case the arguments presented in § 3.1 apply and $v_x \sim a^{-1}$, $v_y \sim b^{-1}$, $v_z \sim c^{-1}$. Under our assumptions, v_y and v_z decrease but $v_x \sim a^{-1} \sim t^{\alpha}$ increases. The perturbation $\Delta\rho$ grows because v_x grows; the motion along the y- and z-axes is unimportant. Let us consider the velocity of one dust particle. The physical velocity is $v_x = a\,dx/dt$. Since $a \sim t^{-\alpha}$ and $v_x \sim t^{\alpha}$, then $dx/dt \sim t^{2\alpha}$ so that $x = x_0 + Dt^{1+2\alpha}$ (here, x_0 and D are constants). In the perturbed solution the volume changes differently from in the unperturbed case: $\Delta V/V \sim t^{1+2\alpha}$, so that $\delta(t) \equiv \Delta\rho/\rho = -\Delta V/V \sim t^{1+2\alpha}$. Keeping in mind that $0 \leq \alpha \leq \frac{1}{3}$, and recalling that $\rho \sim t^{-1}$ for anisotropic expansion, we see that the perturbations grow as $\rho^{-(1+2\alpha)}$, where $1 \leq 1 + 2\alpha \leq 5/3$. Remember that for isotropic expansion, $\delta \sim t^{2/3}$ and $\rho \sim t^{-2}$ so that $\delta \sim \rho^{-1/3}$. The growth rate may be measured by the factor $|d \ln \delta / d \ln \rho|$. For anisotropic expansion this factor is 3 to 5 times larger than for isotropic expansion.

For short waves in an elastic medium (in a relativistic gas with $p = \epsilon/3$, for example, or short-wavelength gravitational waves so that the wavelength satisfies $\lambda \ll ct$) a change in their amplitude is related to a change in the wavelength and frequency by virtue of adiabatic invariance. The wavelength decreases and the frequency increases if the wave propagation takes place along the x-axis. It is moreover apparent that wave propagation along the x-axis, as well as the motion of the dust along the x-axis described in the preceding example, is not an exceptional case. For purely geometrical reasons, an increase in k_x and a decrease in k_y and k_z take place in the course of time even given an arbitrary initial wave vector with components $k_{0x} \approx k_{0y} \approx k_{0z}$ of the same order at $t = t_0$. In other words, the propagation direction of any wave approaches the direction of the x-axis with the passage of time.

Thus, the wave vector is asymptotically directed along the x-axis and grows with time in proportion to t^{α}. Now the speed of the gravitational

wave is constant and equal to c. The speed of elastic waves in the high-temperature plasma (in a gas with $p = \epsilon/3$) is also constant, but equal to $c\sqrt{3}$. In both cases, consequently, the frequency grows in proportion to $k \sim t^{\alpha}$. It follows from adiabatic invariance that the energy of both the gravitational and elastic waves enclosed in a given comoving volume grow as t^{α} too. The comoving volume itself, remember, grows as $\sim t^1$ in the unperturbed Kasner solution.

The energy density of gravitational waves is of order $(dh/dx)^2$, where h is a dimensionless metric perturbation given by the replacement of dx^2 by $(1 + h)dx^2$ in the metric. We thus obtain $h \sim t^{-(1+\alpha)/2}$. The ratio of the energy density of the gravitational waves to the energy density of the plasma grows as $\sim t^{1/3+\alpha}$. In an isotropic model, this ratio would remain constant.[1]

For the elastic (acoustic) waves, $\epsilon = c^2\delta^2$. The relative amplitude of the short waves for which $\lambda \ll ct$ and to which adiabatic invariance applies, grows with time as $\delta \sim t^{\alpha/2}$. In an isotropic model this amplitude is constant. Remember too that the condition $\lambda \ll ct$ also applies to the gravitational waves.

Proceeding by the same methods, i.e., by neglecting the gravitational influence of the matter, one can also easily study the evolution of long-wave perturbations, for which $\lambda \gg ct$, in a medium with $p = \epsilon/3$. Solving the equation of hydrodynamics with respect to the specified metric and neglecting the derivatives of the pressure with respect to ξ, η, and ζ, we find

$$\delta = \delta\epsilon/\epsilon \sim t^{4/3+2\alpha} \sim \epsilon^{-(1+3\alpha/2)} \qquad (20.4.4)$$

for the most rapidly growing density perturbation mode. This replaces the relation $\delta \sim \epsilon^{-1}$ found for the Friedmann model. It is interesting to note that the behavior of the magnitude of the density perturbations does not depend on the equation of state of the matter in the limit of long waves. For both $p = 0$ and $p = \epsilon/3$, we have

$$\rho\delta\,|_{p=0} = \epsilon\delta\,|_{p=1/3} = t^{2\alpha}\,.$$

One should also notice that it is the motion of matter on the background of an unperturbed metric that is under consideration. We cannot describe several types of perturbations, namely, perturbations of the free gravitational field of the model. This is the field that serves as a background for the matter motion, and, as we already have remarked, one can consider this overall field as a gravitational wave of infinite wavelength. It is clear that the adiabatic-invariance method, used above for an analysis of waves satisfying $\lambda \ll ct$, is not applicable to an analysis of gravitational perturbations of large wavelength, $\lambda \gg ct$. Such perturbations of the field as a whole are considered in the analysis by Lifshitz and Khalatnikov (1963*a*, *b*).

1. Remember that it is assumed that radiation plays the dominant role in the plasma, so $p = \epsilon/3$. Also, particle collisions occur so frequently that the isotropy of their distribution and the Pascal law are not destroyed.

Except in special cases, where the symmetry of the problem is important, these perturbations also lead to density and velocity perturbations. However, these modes grow more slowly upon expansion and are not the primary modes.

Other, more subtle questions have not been considered in this section. An example is the conversion of transverse waves into longitudinal waves. For details, we refer the reader to the work done by Doroshkevich (1966).

We now state briefly the main conclusions of this section:

1. During the early stage of an anisotropic cosmological model, the growth of perturbations—matter motions and wave propagation—can be studied without accounting for the self-gravitation of the matter. This is in correspondence with the general conclusion regarding the "vacuum" nature of this stage of the expansion.

2. The law for the growth of small perturbations and the general solution for finite perturbations with zero pressure have been obtained. The perturbations grow more rapidly than in the isotropic model by a factor of 3–5: $|d(\ln \delta)/d(\ln \rho)|_{\text{aniso}} = (3\text{–}5)\,|d(\ln \delta)/d(\ln \rho)|_{\text{isot}}$.

3. The laws describing the amplitude change of short-wavelength gravitational waves and elastic (acoustic) waves have been explained.

4. The law giving the growth of long-wave perturbations in a medium with $p = \epsilon/3$ has been found. The perturbations grow 2–3 times more rapidly in the anisotropic model than in an isotropic model.

20.5 The Instability of the Cosmological Models with Respect to the Emergence of Overall Matter Motions

Until now it has been assumed that the matter is always at rest on the average relative to the reference system of the model, so that $dx^\alpha/dt \equiv 0$. We shall now consider whether such a solution is stable.

In the previous section, the emergence of local fluctuations with $\lambda \ll ct$ was examined. The "kinematic" nature of them was demonstrated in the sense that the self-gravitation of the fluctuations is unimportant. In this section it will be shown that the anisotropic model is unstable in the same approximation, that the fluctuations themselves have no back effects on any aspect of the solution. More precisely, the solution is unstable with respect to the emergence of homogeneous motions of the matter as a whole, i.e., with respect to perturbations with $\lambda \gg ct$.

As for any instability, various modes in this problem grow depending upon whether contraction (collapse) or expansion (the cosmological problem) of the matter as a whole is being considered. For collapse, a similar problem was studied by Lifshitz and Khalatnikov (1960, 1963*a*, *b*). The cosmological problem was examined by Novikov (1970) on the basis of the results obtained by Lifshitz and Khalatnikov. Here the cosmological problem will be studied; about the collapse problem see Landau and Lifshitz (1973).

We shall be considering the vacuum stage, during which the presence of matter does not affect the solution. The metric is then written in the form

$$ds^2 = dt^2 - t^{p_1}(dx)^2 - t^{p_2}(dy)^2 - t^{p_3}(dz)^2 ;$$
$$p_1 + p_2 + p_3 = p_1^2 + p_2^2 + p_3^2 = 1 . \tag{20.5.1}$$

Assume that at some initial instant of time t_0 the matter has a definite but small overall velocity. We shall determine how the matter moves at a time $t \gg t_0$. Since the early stages of expansion are being considered, the equation of state is presumed to be $p = \epsilon/3$. The equation of motion of the matter is then written as (Landau and Lifshitz 1973)

$$\frac{\partial}{\partial x^i}[(\sqrt{-g})\epsilon^{3/4}u^i] = 0 ,$$
$$(p + \epsilon)u^k\left(\frac{\partial u_i}{\partial x^k} - \frac{1}{2}u^l\frac{\partial g_{kl}}{\partial x^i}\right) = \frac{\partial p}{\partial x^i} - u_i u^k\frac{\partial p}{\partial x^k} . \tag{20.5.2}$$

We are here interested in long-wavelength perturbations with $\lambda \gg ct$. In other words, all spatial derivatives with respect to the metric of reference system (20.5.1) are less than the time derivatives. We are thus considering a homogeneous velocity perturbation of the matter as a whole. If the quantities in equation (20.5.2) depend only on time, then the equations are easily integrated. We assume that all physical components of the velocity are of the same order at the initial instant. The expansion of the model is then examined. For $t \gg t_0$, the solution of the system of equations (20.5.2) gives

$$\epsilon = \epsilon_{(0)}t^{-2(1-p_1)} , \qquad u_\alpha = u_{\alpha(0)}t^{(1-p_1)/2} \tag{20.5.3}$$

for the principal terms with respect to powers of t; p_1 is the smallest of the scale-factor exponents. From the solution (20.5.3) and the identity $u_i u^i = 1$, it follows that

$$u_0 = (1 - v^2)^{-1/2} \sim t^{(1-3p_1)/2} .$$

Here v is the modulus of the three-dimensional velocity and is expressed in units of the speed of light. The velocity thus tends to the velocity of light according to the law

$$(1 - v^2)^{1/2} \sim t^{(3p_1-1)/2} . \tag{20.5.4}$$

Let us now find the components of the three-dimensional physical velocity[2] along the various axes. In the given case, these components are

$$v^{(\alpha)} = \frac{u_\alpha(1 - v^2)^{1/2}}{\sqrt{g_{\alpha\alpha}}} . \tag{20.5.5}$$

2. These are the components that a local observer at rest in the reference system measures. In order to avoid a misunderstanding, we should emphasize that this definition does not coincide with that given in Landau and Lifshitz's textbook (1973, see page 412, where the coordinate components of the three-velocity are introduced).

Upon substituting equations (20.5.3) and (20.5.4) and the expression for $g_{\alpha\alpha}$ from the metric (20.5.1), we find from expression (20.5.5) that

$$v^{(x)} \approx \text{const.}\,, \qquad v^{(y)} \sim t^{-(p_2-p_1)}\,, \qquad v^{(z)} \sim t^{-(p_3-p_1)}\,, \qquad (20.5.6)$$

to first order. Thus, $v^{(x)} \gg v^{(y)}, v^{(z)}$, so the velocity tends to that of light along the x-axis, the axis along which contraction of the reference system takes place. One can also write the more exact form $v^{(x)} = 1 - \text{const.}\ t^{3p_1-1}$. Expressions (20.5.3), (20.5.4), and (20.5.6) describe the relativistic stage of motion and demonstrate that the solution is unstable relative to a homogeneous velocity perturbation. The initial stages of development of such an instability, when $v/c \ll 1$, are not described by equation (20.5.3). A description of these stages was obtained in the previous section, however.

We shall now return to our analysis of the relativistic stage of instability and shall turn our attention to the fact that the entire examination has been carried out in the system relative to which the matter moves at a relativistic velocity. This leads to two important consequences.

In the first place, the proper time τ in the comoving system of reference (the system moving together with the matter) differs from the t associated with the metric (20.5.1). The connection between the two times follows from the relativistic relation

$$d\tau = dt(1 - v^2)^{1/2}\,. \qquad (20.5.7)$$

Using equation (20.5.4), we find that

$$\tau \sim t^{(3p_3+1)/2}\,.$$

Substituting this expression into the first of expressions (20.5.3), we then obtain

$$\epsilon \approx \epsilon'_0 \tau^{[-4(1-p_1)/(3p_1+1)]} \qquad (20.5.8)$$

for the energy density.

In the second place, events that are simultaneous in reference system (20.5.1) are not simultaneous in the comoving system. Moreover, with respect to system (20.5.1) there are no spatial gradients of the density and the pressure, for we have chosen a homogeneous solution. However, such spatial inhomogeneity does appear with respect to the comoving reference system because of the relativity of simultaneity.

The presence of pressure gradients in the comoving reference system leads to a dependence of the deformation of the matter on the equation of state. Therefore, materials with different equations of state gather speed relative to system (20.5.1) according to different laws. This is easy to see from equation (20.5.2). For example, for the equation of state $p = 0$, we immediately obtain $u_\alpha = \text{const.}$ from equation (20.5.2). This differs from the result (20.5.3), which holds for the equation of state $p = \epsilon/3$.

For dust, for which $p = 0$, the change in the velocity components of the matter in the system (20.5.1) is a purely kinematic effect; the physical

component of the momentum along this axis is inversely proportional to the scale factor for this axis. The velocity therefore always grows along the axis along which contraction takes place and decreases along the axes along which expansion takes place.

In the presence of pressure, the picture is complicated by the gradients of the pressure forces in the comoving system. Sometimes, the effect of these forces can prevail over the kinematical effect. To demonstrate this, we shall consider the case when the equation of state of the matter is $p = \epsilon/3$ and when the velocity is oriented strictly along one of the axes (call it the α-axis). The remaining components of the velocity are identically equal to zero. Then solution (20.5.3) and relation (20.5.4) are valid. Thus, in this case p_1 can take any value from $-\frac{1}{3}$ to 1 depending upon the axis along which the velocity is directed.

If only the kinematical effect were to play a role in the growth of the velocity, the physical velocity would increase for negative values of p_1 and would decrease for positive values. This actually holds for dust. From equation (20.5.4), however, it is apparent that in the problem under consideration, where $p = \epsilon/3$, the value of the exponent p_1 that is critical for the growth of the velocity is not zero, but rather $\frac{1}{3}$. Thus, even if the motion is directed along the expanding axis, even though H is not expanding extremely rapidly ($p_1 < \frac{1}{3}$), the velocity of the relativistic gas relative to the homogeneous reference system grows. This fact is important for cosmological problems.

Let us return now to the analysis of the original assumptions according to which the problem was examined.

The first remark is related to the assumption that the matter does not react back on the solution for the metric, an assumption equivalent to considering a vacuum stage. This assumption is valid only with the following stipulation. Up to now we have only considered a metric of the form (20.5.1), in which the three-dimensional space is flat, so $\partial g_{\mu\nu}/\partial x^\alpha = 0$ for $\alpha = 1, 2, 3$. For such a metric, $R_0^\alpha \equiv 0$; hence, $T_0^\alpha = 0$ too, since $R_0^\alpha = \kappa T_0^\alpha$. A metric of the form (20.5.1) thus does not admit motion of the matter with respect to the reference system. Strictly speaking, consequently, the existence of matter motions must presuppose a curved three-dimensional space for the homogeneous reference system. If this curvature does not influence the evolution of the model near the $t = 0$ singularity, then it also does not influence the motion of the matter so long as the distance traversed by the matter is small in comparison with the radius of curvature of the three-dimensional space. In this case, the entire account set forth above is valid as a first approximation during the vacuum stage, although it is necessary to remember that the three-dimensional space is curved. We shall see in chapter 22 that for some types of curvature for the three-dimensional space, the $t \to 0$ singularity actually has the Kasner character described above, with $a \sim t^{p_1}$, $b \sim t^{p_2}$, and $c \sim t^{p_3}$, so that the analysis conducted is

valid. For several types of curvature for the three-dimensional space, the curvature changes the character of the solution for $g_{\mu\nu}$ in a radical way near the singularity (Belinsky, Lifshitz, and Khalatnikov 1970). The solution has an oscillatory character and is not described by the simple power-law formulae (20.5.1). Although there is instability in such solutions analogous to that described above, the character of the evolution is, of course, otherwise. These questions are considered in more detail in chapter 22.

The second remark applies to the assumption concerning the homogeneity of the velocity over a very large scale $\lambda \gg ct$. If this condition is fulfilled at some instant for a fixed scale, then it is violated as time passes. Thereafter the terms in equation (20.5.2) involving spatial coordinates play the primary role and change the character of the solution. It is thus always necessary to remember the limitations on the applicability of the solution obtained with respect to future times. Of course, the vacuum solutions are in all cases bounded in time, a matter that was discussed in detail above.

In conclusion, notice the following. We have assumed that, at some initial moment t_0, all components of the velocity are of the same order. The moment t_0 can be the moment when fluctuations in the velocity arise or it can coincide with $t_{\text{Pl}} = 10^{-43}$ s—the minimum time possible in nonquantum cosmology (see Part V). The assumptions made concerning the velocity components then seem reasonable. However, suppose one adopts the point of view that the GTR equations have unlimited applicability, that they apply right to the singularity $t \to 0$. Then, from the assumption that all of the velocity components of the material are nonzero at some instant t_0, it follows that the component along the axis characterized by the most rapid expansion is the largest as $t \to 0$. This axis is the one whose scale factor has the largest positive exponent. Then, near $t = 0$, we approach a picture of relativistic matter motion along the z-axis; the velocity decays in proportion to the expansion factor. We shall not dwell on this[3] in any more detail since GTR is limited with respect to its applicability as $t \to 0$. However, special justification would be needed for such peculiar initial conditions on the velocity (relativistic motion along the z-axis for $t \to 0$). In an examination of a collapsing model (contraction), on the other hand, such behavior for the motion of the matter along the axis with the greatest positive exponent in the time dependence is, of course, necessary.

3. Concerning this, see the result of Lifshitz and Khalatnikov (1963*a*, *b*). Specific models were constructed by Grishchuk, Doroshkevich, and Novikov (1968).

21 THE PHYSICS OF PROCESSES DURING THE EARLY STAGES OF EXPANSION IN ANISOTROPIC MODELS

21.1 Weakly Interacting Particles in an Anisotropic Cosmological Model

Given an anisotropic beginning for the cosmological expansion, the physics of processes during the early stage changes from what it is for isotropic expansion. This is related to two circumstances. First of all, the law for the change in the matter density (and the temperature) with time in anisotropic models differs from the law applying in isotropic models. Second, the role of weakly interacting particles (gravitons and neutrinos) present in the prestellar matter turns out to be extraordinarily important in anisotropic models (Doroshkevich, Zel'dovich, and Novikov 1967*b;* Misner 1967). Accounting for these particles changes sharply the picture of the anisotropic expansion and the physics of processes taking place during the early stage. The reason for the special behavior of the weakly interacting particles is that the various momentum components of freely propagating particles in an anisotropic solution transform differently. An equilibrium distribution of particles, spherically symmetric with respect to momentum space, transforms into a distribution characterized by motion primarily along a single axis (though equal in both directions). It is moreover possible that thermodynamic equilibrium is absent in anisotropic solutions even for superhigh densities. The consequences of these phenomena are discussed in the following sections. Let us immediately note, however, that if an HA solution were actually a correct description of the past history of the Universe, then, as a result of all the processes taking place, it could turn out that the present-day average energy of relic neutrinos is substantially greater than the energy of $T = 2.7$ K relic photons, given, roughly the same average energy density for the two distributions.

Up to now we have passed over questions related to quantum effects at the beginning of the expansion near the singularity, when the spacetime curvature is of the order of $(10^{-33}\ \text{cm})^{-2}$. These effects will be considered in Part V of the book. We are following the papers of Doroshkevich, Zel'dovich, and Novikov (1967*b*, 1969*a*).

We shall now begin a development of the role of weakly interacting particles. The scattering, creation, and annihilation of neutrinos, for ex-

ample, proceeds sufficiently rapidly during the early stages of expansion, when the temperature is high, to maintain thermodynamic equilibrium between the neutrinos and other particles.[1] In particular, the neutrino density varies as $\sim T^3$, the average neutrino energy is then $\sim 3kT$, and the distribution of the neutrinos in momentum space is isotropic. We shall call this stage of expansion the Pascal stage (when the stress-energy tensor is isotropic). Starting at some instant $t = \tau$, however, the neutrino interaction processes become slow by comparison with the expansion rate, and the neutrinos become free particles.

Suppose that in the course of the subsequent expansion ($t > \tau$), the free particles interact neither with other particles nor with one another. We shall see later that this assumption is not always valid for neutrinos and that the picture is more complicated. For gravitons, however, the assumption may always be valid.

In an isotropic solution, the decoupling of the neutrinos does not destroy the equilibrium since neutrinos and photons, as well as e^+e^- pairs, "cool" according to the same law (if $kT > m_e c^2$). In an anisotropic solution, the density and the momentum of free particles changes in correspondence with the cosmological expansion. The momentum components P_x, P_y, and P_z change inversely in proportion to a, b, and c, respectively. During the vacuum stage, $a \sim t^{p_1}$, $b \sim t^{p_2}$, and $c \sim t^{p_3}$, and we shall denote p_1 by $-\alpha$. The energy $E = c|\boldsymbol{P}|$ is essentially determined by the largest momentum component, so the average energy is $\langle E \rangle \sim t^{-p_1} = t^{\alpha}$. The distribution of the particles in momentum space becomes more anisotropic as the age of the Universe increases.

Suppose that at decoupling, $t = \tau$, the energy density ϵ^* of the particles under consideration constitutes a fraction β of the total energy density. In the course of the expansion during the vacuum stage, we then have

$$n^* \sim t^{-1}, \quad E^* \sim t^{\alpha}, \quad \epsilon^* \sim t^{\alpha-1}, \tag{21.1.1}$$

where n^* is the particle number density. For an isotropic stress-energy tensor, the density changes during the vacuum stage according to the law (see eq. [20.1.4])

$$\rho = \frac{3}{32\pi G t^{4/3}\theta^{2/3}} = \frac{K}{t^{4/3}},$$

where $\theta (\theta \geq t)$ is the time at the end of the anisotropic stage of expansion[2]

1. In distinction to isotropic models, anisotropic models allow for the possibility of conditions such that thermodynamic equilibrium is not established (for more detail see § 21.5). Here we are assuming that thermodynamic equilibrium does exist during the early stages.

2. More precisely, θ is defined as the moment of isotropization in a model in which, with all other conditions the same, there are no weakly interacting particles.

and where $K = 3(32\pi G\theta^{1/3})^{-1}$. For free particles, we find from equation (21.1.1) that the law

$$\rho^* = \beta K \tau^{-4/3}(t/\tau)^{\alpha-1} \tag{21.1.2}$$

holds for $t > \tau$. In this case, ρ^* becomes the main component of the total density, and the stress energy tensor is highly anisotropic:

$$-T_0{}^0 = \epsilon \approx \rho^* c^2 \approx T_1{}^1 \gg T_2{}^2, T_3{}^3. \tag{21.1.3}$$

The presence of free particles, whose energy density falls more slowly than does that of interacting particles, markedly decreases the duration of the period when the vacuum solution is applicable. Proceeding by analogy with the determination made in § 20.1, and by using equation (21.1.2), we find that the time $q = t$ after which the vacuum solution fails is

$$q = \theta(\tau/\theta)^{(1+3\alpha)/(3+3\alpha)}\beta^{1/(\alpha+1)}. \tag{21.1.4}$$

The ratio of the energy density of the free particles, ϵ^*, to the energy density of the other particles, $\epsilon - \epsilon^*$, is

$$\frac{\epsilon^*}{\epsilon - \epsilon^*} \approx \frac{\beta}{1-\beta}\left(\frac{q}{\tau}\right)^{\alpha+1/3} = \frac{\beta^{2/[3(1+\alpha)]}}{1-\beta}\left(\frac{\tau}{\theta}\right)^{2(3\alpha+1)/[9(\alpha+1)]} \tag{21.1.5}$$

at the moment when the vacuum solution ceases to be valid. Inasmuch as β is not particularly small, the effects under consideration reduce the duration of the vacuum stage and lead to the maintenance of the relation $\epsilon^* \gg \epsilon - \epsilon^*$ to the end of the vacuum stage.

In a model without free particles after the vacuum stage, rapid isotropization ensues for $t > \theta$. In the model under consideration, there follows after the vacuum stage, for $t > q$, a stage in which the free particles play a dominant role in influencing the dynamics. Moreover, these free particles are described by a sharply anisotropic stress-energy tensor, one satisfying $-T_0{}^0 \approx T_1{}^1 \gg T_2{}^2, T_3{}^3$. A cosmological solution with such a stress-energy tensor is presented in § 20.2. In this solution the gravity due to the directed flux of particles leads to the following asymptotic behavior at large t:

$$\begin{gathered} a \sim t, \quad b \sim (\ln t)^{1/2+p}, \quad c \sim (\ln t)^{1/2-p}; \\ \epsilon^* \sim t^{-2}(\ln t)^{-1/2-2p^2}, \quad n^* \sim t^{-1}(\ln t)^{-p^2-3/4}, \\ E^* \sim t^{-1}(\ln t)^{-p^2+1/4}; \quad 0 < p \leq \tfrac{1}{2}. \end{gathered} \tag{21.1.6}$$

Hence, free particles, whose energy E^* during the vacuum stage grows upon contraction along the x axis, give rise to such a rearrangement of the solution for $t > q$ that their energy begins to fall rapidly.

The change in energy density of these particles with momenta principally along the j-axis is determined by the relation $\epsilon^*{}_j \sim n^* p_j \sim n^* a_j^{-1}$. The change in density of the interacting particles is meanwhile determined by the relation $\epsilon - \epsilon^* \sim n^{4/3}$. Using these relations, one can find that those

noninteracting particles whose momentum is along the x-axis play a dominant role up to the time $t = t_1$, where

$$t_1 \approx \theta\beta^{p-3/2}\delta^{2p/3}[\ln(\beta^{-1}\delta^{-2/3})]^{1/2-p}, \quad \delta = \tau/\theta. \tag{21.1.7}$$

Up to the moment t_1, the energy densities of the particles with momenta principally along the x- and y-axes are equal and strongly exceed the energy density of all other particles $\epsilon^*_x \approx \epsilon^*_y \gg \epsilon^*_z, \epsilon_1$. After t_1, particles with momenta along the y-axis play the main role: $\epsilon^*_y \gg \epsilon^*_x, \epsilon^*_z, \epsilon - \epsilon^*$. Solution (21.1.6) is again applicable but with the scale factor a now characterizing expansion along the y-axis and with other parameter changes. Expansion thus proceeds strongly along the former y-axis but slowly along the former x- and z-axes. Such a rearrangement of the solution gives rise to a rapid fall in ϵ^*_y after a certain amount of time, whereupon a moment is reached when $\epsilon^*_y = \epsilon^*_x$ again ensues. The two axes are hence changing places; the anisotropy is "oscillatory," but with a decreasing amplitude and frequency. This persists until all of the energy densities become of equal order—until such time as $\epsilon^*_x \sim \epsilon^*_y \sim \epsilon^*_z \sim \epsilon - \epsilon^*$. We are assuming here that β is of order unity. After this time, isotropization of the solution ensues.

To estimate the isotropization time t_2, we consider two limiting cases.

1. In solution (21.1.6), the y- and z-axes are equivalent, i.e., suppose that $p = 0$. It then turns out that the $\epsilon^* \approx \epsilon - \epsilon^*$ up to the end of the first period of "oscillations," when $t = t_1$. After this, rapid isotropization ensues. For values of α close to $\frac{1}{3}$, consequently, the isotropization time t_2 obtained from equation (21.1.7) for $p = 0$ is

$$t_2 \approx \theta\beta^{-3/2}[\ln(\beta^{-1}\delta^{-2/3})]^{1/2}. \tag{21.1.8}$$

2. The other limiting case corresponds to $\alpha = 0$. In this case

$$\epsilon^*_x \approx \epsilon^*_y = \text{const.}, \quad \epsilon^*_z \sim t^{-2}, \quad \epsilon - \epsilon^* \sim t^{-4/3} \tag{21.1.9}$$

during the vacuum stage.

After the vacuum stage ends, it is not particles whose momentum is along any one axis that play the dominant role (as in eqs. [21.1.6]), but rather particles in the x-y plane whose momentum may, with equal probability, be in any direction in this plane. The stress-energy tensor for the particles is then

$$-T_0{}^0 = \epsilon, \quad T_1{}^1 = T_2{}^2 = \frac{\epsilon}{2} \gg T_3{}^3. \tag{21.1.10}$$

By neglecting the energy of all other particles, we obtain the exact solution

$$a = b = a_0\left(1 + \frac{t}{t_0}\right)^{2/3}, \quad c = c_0\left[1 - \left(1 + \frac{t}{t_0}\right)^{-1/3}\right];$$
$$\kappa\epsilon = \tfrac{4}{9}t_0^{-2}\left(1 + \frac{t}{t_0}\right)^{-5/3}\left[\left(1 + \frac{t}{t_0}\right)^{1/3} - 1\right]^{-1}. \tag{21.1.11}$$

The effective time for a substantial change in the vacuum solution is t_0. From equation (21.1.9) it follows that $\epsilon^*_x = \epsilon^*_y \gg \epsilon^*_z$, $\epsilon - \epsilon^*$ holds until the time $t_0 \approx \beta^{-1}\theta\delta^{1/3}$.

In analogy with solution (20.2.9), the expansion for $t > t_0$ again proceeds in such a way that the energy density of particles with momenta in the x–y plane falls more rapidly than do ϵ_z and ϵ_1. At the time

$$t_1 \approx \beta^{-5/2}\theta\delta^{-2/3}\,, \tag{21.1.12}$$

ϵ^*_x, $\epsilon^*_y \approx \epsilon^*_z \approx \epsilon_1$ (β is of order unity); isotropization thereupon ensues.

Thus, for α close to zero, when more rapid expansion proceeds sometimes along the x-axis and sometimes along the y-axis, one can make use of the relation ϵ^*_x, $\epsilon^*_y \gg \epsilon^*_z$, $\epsilon - \epsilon^*$ for an estimate of the isotropization of time. In this case, consequently,

$$t_2 \approx \theta\delta^{-2/3}\,. \tag{21.1.13}$$

In the general case, the isotropization time lies between the values (21.1.8) and (21.1.13). In addition, isotropization of the solution in the general case proceeds in an oscillatory fashion, so that when the anisotropy finally becomes small, the solution is of the form

$$a_j \sim a_{j0}t^{1/2}\left\{1 + C_j t^{-1/4}\sin\left[\frac{\sqrt{3}}{2}\ln(D_j t)\right] + \ldots\right\}\,, \tag{21.1.14}$$

where a_j represents a, b, or c, and where C_j and D_j are constants. The isotropic expansion of free particles is hence stable.

Notice finally that the decoupling of gravitons takes place under conditions such that the quantum effects of gravitation (see 7.2) are substantial. At the time of decoupling, therefore, the gravitons may also not be found in equilibrium with the other particles. Given equilibrium, $\beta \sim 10^{-2}$; if equilibrium does not exist, then β may be even further from unity. If $\beta \ll 1$, the relation $\epsilon_G \approx \epsilon - \epsilon_G$ must still hold after isotropization; i.e., the present-day graviton energy density must be of the order of the photon energy density, which is roughly 10^{-12} ergs cm^{-3} or 10^{-33} g cm^{-3}. In this case, for $\beta \ll 1$, the momentum distribution of the gravitons is anisotropic. Hence, if at the instant of decoupling their average energy is of the order of the energy of the other particles but their density is small, $n_G \ll n_1$, then $n_G \ll n_1$ and $\langle E_G\rangle \gg E_\gamma$ will also hold today.

Up to now we have not considered the possibility that the processes described are unstable. The anisotropic vacuum solution itself becomes unstable with a power-law growth (Lifshitz and Khalatnikov 1963*a*, *b*); but if the initial perturbations are small, then this instability does not manage to manifest itself. However, isotropization processes are possible that involve a precisely oriented flux of particles, a flux that arises due to a collective interaction. We shall not consider this here, but notice only that if such an isotropization process occurs during the stage when $E_G \gg$

E_γ, then the present-day energy of the gravitons is greater than the energy of the RR photons.

21.2 Neutrinos in an Anisotropic Model

The true behavior of neutrinos differs from the behavior just studied in the following way. Given anisotropy, with contraction along one of the axes during an early stage, some of the neutrinos and the oppositely moving antineutrinos gain such an energy that, under certain conditions, there develops a finite probability for their irreversible conversion into electrons and positrons.

We shall at first suppose that there are no gravitons at all[3] and shall consider for definiteness the behavior of electron neutrinos. We shall examine the joint behavior of the neutrinos and gravitons later.

We find the neutrino decoupling time τ from the condition of equality of the hydrodynamical and relaxation times, which gives

$$\sigma n c \tau = 1 . \tag{21.2.1}$$

Here σ is the cross section for the interaction $e^+ + e^- \rightleftarrows \nu_e + \bar{\nu}_e$ and $\sigma \sim E^2$ if $E < 300$ GeV. (We shall assume that this condition is fulfilled; for the case when $E > 300$ GeV, see the end of this section; according to some of the modern theories the upper limit is 35 GeV.) Furthermore, n is the particle density and c is the speed of light.

We shall suppose that $\tau < \theta$ since, in the opposite case, the neutrinos decouple after isotropization of the solution and there are no anisotropy effects. During the Pascal stage before the time τ, $\sigma \sim E^2 \sim t^{-2/3}$ and $n \sim t^{-1}$. In an isotropic Friedmann model, recall that $\sigma \sim E^2 \sim t^{-1}$ and $n \sim t^{-3/2}$. Knowing that, for the isotropic case, the decoupling time is $\tau' \approx 0.1$ s for ν_e (the decoupling temperature is ~ 3 MeV), and using the relations already written, we shall express τ in terms of τ' and θ. We obtain

$$\tau = (\tau')^{9/4} \theta^{-5/4} . \tag{21.2.2}$$

The decoupling time is determined by the energy density and the rate of the bulk expansion, and, since these quantities do not depend on α, τ also does not depend on α.

For $t > \tau$ the average energy of the neutrinos with momenta along the x-axis grows as $\langle E_\nu \rangle \sim p_x \sim t^\alpha$, where $0 \leq \alpha \leq \frac{1}{3}$. The interaction cross section increases correspondingly. For the neutrino scattering process

$$\nu_e + e \rightleftarrows \nu_e + e , \tag{21.2.3}$$

the cross section is proportional to the square of the energy of the ν_e and the e in the center of inertia system: $\sigma \sim E_{ei}^2$. For relativistic electrons

3. More precisely, we shall assume that the energy density of the gravitons is much less than ϵ_i up to the moment of the neutrino decoupling.

in the laboratory system, we have $\sigma \sim E_\nu E_e$. The condition that the process play a role for the neutrinos is $w \equiv \sigma n c t > 1$; at decoupling $w = 1$. How does this quantity change during the subsequent evolution? Notice first that the dependence of the individual quantities on time is

$$E_\nu \sim t^\alpha , \quad \alpha < 1/3 ; \quad E_e \sim t^{-1/3} ; \quad n_e \sim t^{-1} .$$

Since $\sigma \sim E_e E_\nu$, w necessarily decreases; $w < 1$ for $t > \tau$. By itself this process does not lead to a violation of the condition $\sigma n c t < 1$ and to a noticeable transfer of the energy of the free neutrinos into that of e^+e^- pairs (concerning the significance of this process given neutrino annihilation, see the following paragraph).

The neutrinos and antineutrinos moving oppositely along the x-axis do, however, interact. In fact, the cross section for the annihilation process,

$$\nu_e + \bar{\nu}_e \to e^+ + e^- , \tag{21.2.4}$$

satisfies $\sigma \sim E_\nu^2 \sim t^{2\alpha}$, so, given conservation of the number of particles, the condition $\sigma n c t < 1$ is violated. During the vacuum stage, process (21.2.4) leads to an irreversible conversion of neutrinos into e^+e^- pairs that are instantaneously thermalized. Let us now find the rate of growth of the average neutrino energy $\langle E_\nu \rangle$ and the decrease in their density n_ν under these conditions.

In momentum space, the neutrino distribution is represented by an ellipsoid with axes p_x, p_y, and p_z. At each instant, process (21.2.4) limits the magnitude $|p_x|$ to a certain value. For the remaining particles, with $|p_x|$ less than this value, the Liouville theorem is valid; therefore $n_\nu \sim |p_x p_y p_z|$, while the cross section satisfies $\sigma \sim p_x^2$. From the condition $\sigma n c t = 1$, we then find that

$$E_\nu \sim t^{\alpha/3} , \quad n_\nu \sim t^{-(1+2\alpha/3)} , \quad \epsilon_\nu \sim t^{-(1+\alpha/3)} . \tag{21.2.5}$$

From the law of energy conservation, it is not difficult to calculate the rate of change of the total energy density $\epsilon_{\gamma,e^\pm}$ of e^+e^- pairs and photons in equilibrium when account is taken of the "heating" by process (21.2.4). It is also not difficult to calculate the rate of entropy growth due to such heating. For $t \gg \tau$, we obtain

$$\frac{\epsilon_{\gamma,e^\pm}}{\epsilon_\nu} \approx \frac{4\alpha}{1-\alpha} + \left(\frac{\tau}{t}\right)^{(1-\alpha)/3} \tag{21.2.6}$$

and

$$\frac{S}{S_0} \approx \left[1 + \frac{4\alpha}{1-\alpha}\left(\frac{t}{\tau}\right)^{(1+\alpha)/3}\right]^{3/4} , \tag{21.2.7}$$

respectively. The time when the vacuum stage comes to an end, $t = q$, is found just as in the preceding sections. It is (note that by comparison with eq. [21.1.4] the exponent has changed somewhat and the isotropization time has increased)

$$q \approx \theta(\tau'/\theta)^{9(1-\alpha)/[4(3-\alpha)]} , \quad \tau' < \theta . \tag{21.2.8}$$

Up to the time $t = q$, the ratio of the number densities and the ratio of the average energies satisfy

$$\frac{n_\nu}{n_{\gamma,e^\pm}} \approx \frac{\langle E_{\gamma,e^\pm}\rangle}{\langle E_\nu\rangle} \approx \left(\frac{\tau'}{\theta}\right)^{3(3+5\alpha)/[8(3-\alpha)]} \qquad (21.2.9)$$

for large values $\alpha > (\tau/q)^{(1-\alpha)/3}$. After $t = q$, the expansion proceeds in all directions, and the interaction between the neutrinos and the anti neutrinos ceases. For $t > q$,

$$S(t) \approx S(q)\ ; \quad \frac{n_\nu}{n_{\gamma,e^\pm}} \approx \frac{n_\nu}{n_{\gamma,e^\pm}}\Big|_{t=q}\ ; \quad \frac{E_\nu}{E_{\gamma,e^\pm}} \approx \frac{E_\nu}{E_{\gamma,e^\pm}}\Big|_{t=q}.$$

For α not small, it follows from equation (21.2.6) that the energy densities of the neutrinos, the photons, and the pairs are of the same order up to the time $t = q$. The large anisotropy rapidly disappears in an isotropization time of order q. Inasmuch as $n_\nu \ll n_{\gamma,e^\pm}$ (see eq. [21.2.9]), however, $\langle E_\nu\rangle \gg \langle E_{\gamma,e^\pm}\rangle$ and the distribution of the neutrino momenta is sharply anisotropic.

In the relations given in this section, the factor β is not taken into account. For neutrinos, it can assume values (in its dependence on τ) from 10^{-2} to $\frac{1}{4}$.

Earlier in this section it was assumed that the condition $\sigma_\nu \sim E_\nu^2$ on the neutrino annihilation cross section (see eq. [21.2.4]) is valid. However, this condition is apparently fulfilled only for $E_\nu \lesssim 300$ GeV. For larger energies, the cross section for neutrino annihilation either remains constant or even decreases with increasing energy. The results obtained above are therefore totally valid only for $q \leq 10^4$ s ($\theta \leq 3 \times 10^{10}$ s, $\alpha = \frac{1}{3}$). For $q = 10^4$ s, one obtains $\theta \approx 3 \times 10^{10}$ s, $\tau = 10^{-15}$ s, $E_\nu|_{t=\tau} \approx 3$ GeV, $E_\nu|_{t=q} = 300$ GeV, and $S(q) \approx 10^3 S(\tau)$. Under these conditions, the present-day neutrino energy is $E_\nu \approx 10^4$ eV. At high energies, the cross section depends on assumptions concerning the existence of and the masses of the so-called intermediate, charged mesons ($W^\pm$) (see Okun' 1966).

For $\theta > 3 \times 10^{10}$ s ($E_\nu|_{t=\tau} > 3$ GeV), the oppositely moving neutrinos and antineutrinos effectively annihilate into e^+e^- pairs only so long as $E_\nu \leq 300$ GeV. For larger energies, the growth in the cross section slows. The larger the value of θ, the earlier the time when the energy of the neutrinos achieves the value 300 GeV and the earlier the time when the growth in the annihilation of the neutrinos and the growth in the entropy slows. We denote the former time by t_3. For $\theta > 3 \times 10^{10}$ s, the formulae in this section are only applicable for times satisfying

$$t \leq t_3 \approx 10^{15/\alpha}\tau' \left(\frac{\tau'}{\theta}\right)^{(3+5\alpha)/4\alpha}, \qquad \alpha \sim \tfrac{1}{3}\ .$$

For $t \geq t_3$ it is necessary to make use again of the formulae of § 21.1. Upon an increase in θ (for $\theta > 3 \times 10^{10}$ s), the isotropization time is increased,

while the growth in the entropy in the course of the deformation is decreased, as is the final energy of the present-day neutrinos. The parameter θ cannot be larger than about 10^{14} s since, in the opposite case, the expansion could not be isotropized by the present time and would give a large RR anisotropy. Estimates show that the maximum increase in the entropy takes place for the case described as follows:

$$\theta = 10^{19}\ \text{s}, \quad \alpha \approx 0.02, \quad \tau = 10^{-26}\ \text{s}, \quad q \approx 200\ \text{s} \cdot$$

$$\frac{S(q)}{S(\tau)} \approx 10^6; \quad \frac{n_\nu}{n_\gamma} \approx 3 \times 10^{-8}.$$

Here n_γ is the RR photon number density. The energy of the neutrinos is, in this case, $E_\nu \leq 3 \times 10^4$ eV today. Now within the realm of the scheme under consideration, $S(q)/S(\tau) \leq 10^6$ always holds and the neutrino energy is $E_\nu \leq 3 \times 10^4$ eV today. However, one ought to keep in mind that $T_e \sim t^{-(3+\alpha)/12}$ for the conditions under consideration. For process (21.2.3), therefore, just as for the annihilation process, $\sigma n t = \text{const}$. As a consequence, process (21.2.3) also contributes to the entropy gain. Inasmuch as the annihilation process remains the primary one, however, the contribution of the scattering process (21.2.3) cannot substantially change the estimates presented above. If, by the time of neutrino decoupling, the graviton energy density becomes much greater than ϵ_1, then the kinematics of the expansion are determined by the gravitons and the processes involving neutrinos are determined in turn by the kinematics.

Let us now summarize briefly what has been stated in this section.

1. The nature of the dynamics of anisotropic cosmological models and of the physics of processes in them is closely connected with the presence of weakly interacting particles and with the possible nonequilibrium character of the matter present.

2. The average energy of the weakly interacting particles (neutrinos and gravitons) at the present time can differ sharply from that predicted by an isotropic model and can be quite large, given a corresponding decrease in their number per unit volume.

3. A large increase in the initial entropy of the matter is possible in the anisotropic models.

21.3 The Effect of Viscosity on the Expansion of Anisotropic Models

In Misner's work (1967, 1968), the questions concerning the behavior of weakly interacting particles in anisotropic cosmological models have been approached differently from the analysis set forth in the previous section. Misner realized that one can describe small deviations from an equilibrium particle distribution function by means of an increase in the viscosity and, also, that viscosity leads to a growth in entropy. This growth

is greater, the more the distribution function differs from an equilibrium form and the greater the relaxation time. However, one can make use of the concept of viscosity only so long as the relaxation time is less than the hydrodynamical time. If we remove this restriction, Misner's calculations lead to the conclusion that, because of the effect of viscosity, the entropy gain is always so significant that the neutrinos do not become free during the vacuum stage. According to Misner, therefore, the regime examined in the preceding section does not occur and, irrespective of the parameters of the model, the isotropization takes place at temperature $T \approx 3$ MeV. Only after the isotropization, according to Misner, do the neutrinos become free, so that there are no high-energy neutrinos today.

In this section we shall briefly examine this viscosity effect, relying principally on Misner's ideas. In the next section, we shall set forth the kinematic equation that arises in describing the theory of neutrinos in an anisotropic cosmology. On the basis of this more general approach, it will be determined when the viscosity approximation is valid and when the free-particle approximation is valid.

With due regard for the viscosity, the equation for the conservation of energy in a model with the Kasner metric (19.3.6), (19.3.7) takes the form

$$\frac{d\epsilon}{dt} + \frac{4}{3}\frac{\epsilon}{l} = \frac{4}{3}\frac{\eta}{l^2}, \tag{21.3.1}$$

for the vacuum stage. The quantity η is the coefficient of viscosity. We shall take

$$\eta = \tfrac{1}{3} t^* \epsilon_\nu = \frac{k}{3}\frac{\epsilon}{\sigma n_e}. \tag{21.3.2}$$

Here t^* is the mean free time and $\sigma \sim (E_\nu E_e)^m \sim T^{2m}$ the cross section for the scattering of ν from $e^\pm$ $(m > 0)$; n_e and E_e are the number density and energy, respectively, of the scatterers (electrons and positrons); and $k = \epsilon_\nu/\epsilon$ can, given equilibrium, be expressed in terms of the statistical weights of the neutrinos and of the remaining particles. Note that, for equilibrium, k coincides with β introduced in § 21.1.

As with Misner, we assume that k = const., i.e., that the number of kinds of particles remains unchanged over the course of the entire process.

The solution of equations (21.3.1) and (21.3.2) is easily obtained in the form

$$\epsilon = \epsilon_{ad}(t)[1 + Ct^{2m/3}]^{4/(3+2m)},$$

$$\epsilon = \epsilon_{ad}(t)\left[1 - k\left(\frac{3+2m}{6m}\right)\left(\frac{t^*}{t}\right)\right]^{-4/(3+2m)}, \tag{21.3.3}$$

where $\epsilon_{ad}(t) \sim t^{-4/3}$ follows from the adiabatic law for the change in the energy density. Equation (21.3.3) then leads, for $Ct^{2m/3} \gg 1$, to

$$\epsilon \sim t^{-4/(3+2m)}, \tag{21.3.4}$$

independent of the initial parameters of the problem.

Misner considered the case for $m = 0.5$ and $k = 1$ and found that $\epsilon \sim t^{-1}$ during the vacuum stage and that the neutrinos are in equilibrium up to the isotropization time. Irrespective of the initial parameters, the isotropization occurs when $T \approx 3$ MeV.

In reality, the main question concerns the grounds for the approximation leading to equation (21.3.4). It follows from equation (21.3.3) that the regime under consideration is established in the limit

$$k[(3 + 2m)/(6m)]\ (t^*/t) \to 1.$$

Given real conditions, however, many particles are in thermodynamic equilibrium. For $T \sim 0.5$ GeV, e, γ ν_e, μ, ν_μ, and π are in equilibrium, while for $T > 3$ GeV, it is also necessary to take account of baryons and of scalar and vector mesons.

In its dependence on temperature, therefore, $k = \epsilon_\nu/\epsilon$ lies within the interval $0.25 \geq k \geq 0.025$. Consequently, in its dependence on the value of k, the solution gives the regime under consideration, that examined by Misner, $t^*/t = (4.8\text{–}48)$ if $m = 1$, or $t^*/t = (3\text{–}30)$ if $m = 0.5$. But these are conditions under which the applicability of the concept of viscosity is certainly not evident. Nevertheless, the inadmissibility of the macroscopic description of the phenomenon by means of the concept of viscosity still does not mean that Misner's conclusions are qualitatively incorrect, since it is possible that a solution of the kinetic equation exists with properties similar to the properties of the solution obtained by Misner. To obtain reliable results under these conditions, it is therefore necessary to study the kinetic equation. At the same time, though, the additional possibility then presents itself that the processes described in the previous section are realized. One cannot then talk about viscosity. So, with the goal of reliable results in mind, let us proceed to a study of the kinetic equation.

21.4 The Kinetic Theory of Neutrinos in an Anisotropic Model; the Self-similar Solution

The kinetic theory of neutrinos in an anisotropic model was first investigated by Doroshkevich, Zel'dovich, and Novikov (1968, 1969*a*, *b*). We shall be considering the vacuum stage and shall moreover suppose that $\alpha = \frac{1}{3}$, i.e., that $a \sim t^{-1/3}$ and $b = c \sim t^{2/3}$. The momentum-space distribution function for the neutrinos (with density w), referred to a cell of volume $(2\pi\hbar)^{-3}$, is determined by the kinetic equation

$$\frac{\partial w}{\partial t} + \frac{p_x}{3t}\frac{\partial w}{\partial p_x} - \frac{2}{3}\frac{p_y}{t}\frac{\partial w}{\partial p_y} - \frac{2}{3}\frac{p_z}{t}\frac{\partial w}{\partial p_z} = \left(\frac{\partial w}{\partial t}\right)_{\text{coll}}. \qquad (21.4.1)$$

Here $(\partial w/\partial t)_{\text{coll}}$ is the collisional term, which describes the change in the neutrino distribution function resulting from interactions with the other particles.

The other particles are in a state of thermodynamic equilibrium char-

acterized by a specified temperature T (in energy units). We shall assume that all of these particles are ultrarelativistic so that the equation of state $P = \epsilon'/3$ holds for them. Their energy density is then $\epsilon' = \kappa T^4$, with $\kappa = b24\pi/(2\pi\hbar)^3$ and where b is the sum of the statistical weights. We shall neglect the departure from classical behavior of the Fermi and Bose distributions. That is, we shall replace $n^* \equiv (e^{E/T} \pm 1)^{-1}$ by $e^{-E/T}$ in formulae describing the equilibrium distribution and shall correspondingly discard n^* in the factors $(1 - n^*)$ and $(1 + n^*)$ in the collision integral. Then $b = \Sigma g_j$, where g_j is the statistical weight of the jth kind of particle ($g = 2$ for $e^{\pm}$, $g = 1$ for the scalar mesons, and $g = 1$ for the two-component neutrinos). Given these assumptions,

$$k = \frac{g_{\nu\bar{\nu}}}{b + g_{\nu\bar{\nu}}} = \frac{2}{b + 2} .$$

The temperature is determined by the conservation-of-energy equation

$$\frac{d\epsilon'}{dt} + \frac{4}{3}\frac{\epsilon'}{t} \equiv 4\kappa T^3 \left(\frac{dT}{dt} + \frac{T}{3t}\right) = Q(W, T) , \qquad (21.4.2)$$

where $Q(W, T)$ is the term describing the effect on the temperature of the interaction of the plasma with the neutrinos. If one assumes that the cross section for the interaction of the neutrinos with the other particles has a power-law dependence on the energy of the colliding particles in the center-of-energy system, $\sigma = \sigma_0 E^{2m}$, then one can convince oneself that the system of equations (21.4.1) and (21.4.2) admits a self-similar solution of the form

$$w = w(q) , \quad T \sim t^{-1/(2m+3)} , \qquad (21.4.3)$$

where

$$q = p/T .$$

The physical parameters defining the problem for a given metric are the exponent m, the constant σ_0 (or several constants σ_{0j}) in the expression for the interaction cross section, and the numbers b and k characterizing the thermodynamics of the system. To determine the region where the self-similar solution exists, it is convenient to do as follows: after specifying the values of σ_0 and $T(t)$, then find the region over which k changes.

One can successfully carry out this general program only after making additional simplifying assumptions relating to the interaction of the neutrinos with the equilibrium plasma and with one another. The problem can be successfully solved in two cases.

A. In the first case, only the isolated birth and death of single neutrinos via the reactions

$$\nu + n \leftrightarrow p + e \qquad (21.4.4)$$

is considered. Here, instead of n, p, and e, other particles found in the equilibrium stage can also play a role. Besides them, other particles not interacting with the neutrinos can also be present.

B. In the second case, only the pairwise annihilation and creation of neutrinos and antineutrinos according to the reactions

$$\nu + \bar{\nu} \rightleftharpoons e^+ + e^- \tag{21.4.5}$$

are considered. There is here an additional limitation, $m = 1$, on the exponent in the expression for the cross section. Here too, other particles can be present that do not interact with the neutrinos.

An account of the scattering of the neutrinos and a consideration of the second case without the limitation $m = 1$ lead to great technical difficulties, so such variants have still not been studied. In the problem with scattering only (without creation—either pairwise or single), there is evidently no self-similar solution.

Let us now summarize briefly the results obtained above. For details see Doroshkevich, Zel'dovich, and Novikov (1969*a*).

1. The self-similar solution of the type considered above is possible for $m = 1$ only if $(1 - k)/k \ll 1$. This follows from the study of the second case as well as from a study of the first case. Apparently, an account of the scattering process $e + \nu \leftrightarrow e + \nu$ cannot noticeably change the result.

2. The existence of critical values of the exponent, $m_0 \approx 0.545$ and $m_1 \approx 1.5$ follows from an analysis of the first problem. These values are such that for $m < m_0$, the self-similar solution exists for any value of k satisfying $0 < k < 1$. For $m_0 < m < m_1$, the self-similar solution exists only for $1 \geq k \geq k_{\min}$. Finally, for $m > m_1$, there is no self-similar solution at all. For the general solution of the kinetic equations, see Doroshkevich, Zel'dovich, and Novikov (1969*a*, *b*).

From what has been said, one can draw the following conclusions. For the most important (realistic) value, $m = 1$, i.e., for $\sigma \sim E^2$, and for a realistic value of $k = \epsilon_\nu/\epsilon < 0.25$, with $\alpha = \frac{1}{3}$, there is no self-similar solution of the kinetic equation similar to Misner's solution. Such a solution is obtained upon a formal application of the equations of hydrodynamics with viscosity for the values $t^*/t > 1$. This is evidently unacceptable and even leads, as we see from an analysis of the kinetic equation, to a qualitatively incorrect result.

The physical conclusion is that when deviations from equilibrium begin for the weakly interacting particles in an anisotropic cosmology with α close to $\frac{1}{3}$, then the number of particles decreases soon after they become free. Their average energy, however, grows. This was described in § 21.2. Subsequent work by Matzner and Misner (1972) confirms this conclusion.

21.5 Primordial Nucleosynthesis in Anisotropic Models

We shall turn now to an analysis of the effect of anisotropy on nuclear reactions and other particle processes. What is important there is that the expansion occurs differently from the isotropic case.

As was noted in chapter 19, expansion from the singular space proceeds in a very simple case according to the Kasner solution. With matter at rest, the volume in this solution varies as $V \sim t$ and the energy density as $\epsilon \sim t^{-4/3}$. In a more general case, when the matter moves at an ultrarelativistic velocity relative to the homogeneous reference system, the density changes as

$$\epsilon = \epsilon'_0 \tau^{-4(1-p_1)/(3p_1+1)} \tag{21.5.1}$$

in its dependence on the proper time τ. This proper time does not coincide with the time t of the homogeneous model since the matter is moving in the reference system at an ultrarelativistic velocity. One can rewrite this expression in the form

$$\epsilon = \epsilon'_0 \tau^{-\beta}; \quad \beta = \frac{4(1-p_1)}{3p_1+1}. \tag{21.5.2}$$

It was noted (§ 20.5) at the end that given a uniform selection of initial conditions, the quantity β can run through all of the values from 0 to ∞. However, given the most natural (see § 20.5) choice of initial conditions, the exponent β varies from 4 to ∞. Of course, after the first vacuum stage of expansion in an anisotropic model, other stages with a different expansion law follow. While keeping this in mind, we shall not be specifying all of the possible values of β here but shall examine how the physical processes in the expanding matter proceed for the entire spectrum of values, $0 \leq \beta < \infty$.

We begin with an examination of the question of the nature of the thermodynamic equilibrium during the early stages of expansion. Complete equilibrium exists if the characteristic time t^* for reactions working to establish equilibrium is less than the hydrodynamical time, the time from the beginning of the expansion:

$$t^* = \frac{1}{\sigma n} < \tau, \quad \sigma n \tau > 1. \tag{21.5.3}$$

Here n is the particle number density, σ is an appropriate interaction cross section, and we have set $c = 1$.

In the Friedmann model with $p = \epsilon/3$, the quantity n satisfies $n \sim V^{-1} \sim \tau^{-3/2} \sim T^3$, where T is the temperature. Therefore, $n\tau \sim \tau^{-1/2} \sim T$. Consequently, if σ grows with energy, or in any event falls more slowly than T^{-1}, then the left-hand side of inequality (21.5.3) tends to infinity for $\tau \to 0$, where $T \to \infty$. During the early stages of the Friedmann model, therefore, complete thermodynamic equilibrium exists for all particles except, perhaps, gravitons (see Part II).

During the Kasner stage of expansion, $n\tau = \text{const.}$ as $\tau \to 0$ for matter at rest. Even in this case, if σ tends to a constant or falls as the energy grows, then it is possible that, when $\tau \to 0$, equilibrium among several of the types of particles or even among all of the types does not exist.

In the general case when $\sigma = \sigma_0 T^{\beta}$, we have

$$n\tau \sim \tau^{1-3\beta/4} . \tag{21.5.4}$$

If $\beta < 4/3$, then the product $n\tau \to 0$ as $\tau \to 0$. In this case, there is during the early stages of expansion no thermodynamic equilibrium, and the particles are free (provided the cross section σ does not tend to infinity as $T \to \infty$, faster than $\sigma \sim T^{(4/\beta-3)}$). If there is no thermodynamic equilibrium, then, as a consequence of the deformation anisotropy, the pressure is anisotropic and it is necessary to consider the effects described above in § 21.1. For the most realistic case of $\beta > 4$, however, equilibrium holds even if the cross section σ tends to zero as $T \to \infty$ (but not faster than $\sigma \sim T^{4/\beta-3}$ or, for very large values of β, not faster than T^{-3}).

We shall now make some brief remarks concerning the possibility that thermodynamic equilibrium is absent and shall then pass on to the problem of nucleosynthesis in anisotropic cosmologies. If thermodynamic equilibrium does not exist among the particles during the early stage of expansion, then one evidently cannot say anything definite about the results of nuclear reactions and the abundances of the elements in the prestellar matter. Everything then depends on the arbitrary choice of the "initial" composition of the matter. Having said this, we shall, however, suppose that thermodynamic equilibrium among the known particles does exist (up to the aforementioned [see §§ 21.2–4] destruction of the equilibrium due to the neutrinos).

We now repeat briefly the description of the processes of element synthesis given in Part III. In the hot isotropic, Friedmann model, nucleosynthesis proceeds in two stages. During the early stage, equilibrium between the neutrons and the protons is maintained by the weak interactions, for example reactions such as

$$e^+ + n \rightleftarrows p + \bar{\nu} , \qquad \nu + n \rightleftarrows p + e^- . \tag{21.5.5}$$

The characteristic time for these reactions at high temperature is (where $T_9 \equiv T/10^9$ K)

$$\tau_{\text{weak}} \approx \frac{10^5}{T_9^5} \text{ s} . \tag{21.5.6}$$

The ratio of the number of neutrons to protons in equilibrium is

$$\left(\frac{n_n}{n_p}\right)_{\text{equil}} = \exp\left(-\frac{\Delta m c^2}{kT}\right) \approx \exp\left(-15/T_9\right) , \tag{21.5.7}$$

where Δm is the proton-neutron mass difference: $m = m_n - m_p$. The change in the temperature in the course of the expansion is given as

$$\tau_{\text{exp}} \approx \frac{10^2}{T_9^2} \text{ s} . \tag{21.5.8}$$

As we already noted in chapter 6, the time required for a substantial change in the temperature in an expanding model is of the order of τ_{exp}.

For thermodynamic equilibrium to hold, it is necessary that $\tau_{weak} \ll \tau_{exp}$. At the beginning of the expansion thermodynamic equilibrium is established since, for $T_9 > 10$, $\tau_{weak} \ll \tau_{exp}$. When τ_{weak} becomes greater than τ_{exp}, thermodynamic equilibrium between n and p is destroyed; reactions (21.5.5) thereafter cannot proceed sufficiently rapidly and the ratio n_n/n_p "freezes out." By equating equations (21.5.6) and (21.5.8), we find for the temperature at "freeze-out," $T_9 = 10$. Substituting $T_9 = 10$ into equation (21.5.7), we then obtain

$$\left(\frac{n_n}{n_p}\right)_{freeze} \approx e^{-1.5} \approx 0.2 . \tag{21.5.9}$$

In the second stage of nucleosynthesis, nuclear reactions occur in the course of the subsequent expansion, at lower temperatures, and the formation of the nuclei of light elements becomes possible. A large fraction of the neutrons are captured by protons, so that mainly ^{4}He and some very small amounts of ^{3}He, Li, and D result. If all of the neutrons are captured, then the ^{4}He abundance by mass is

$$Y_{max} = \left(\frac{2n_n}{n_n + n_p}\right)_{freeze} \approx 0.33 . \tag{21.5.10}$$

Actually, the true freeze-out ratio n_n/n_p (see Fig. 31) is less than that obtained from the estimate of n_n/n_p when $\tau_{exp} = \tau_{weak}$. In addition, not all neutrons are captured before they decay. As a result, the calculated helium abundance is less than the estimate in equation (21.5.10), $Y \lesssim 0.3$. Nevertheless, the estimate (21.5.10) is quite satisfactory.

In HA models the rate of change of the density of the matter with time is different from what it is in an isotropic model. This leads to different results for the nuclear reactions that proceed in the expanding matter and, as a consequence, to a different chemical composition for the primordial matter. Consider as an example the very simple Heckmann-Schücking anisotropic model with a flat comoving space such that it gives the present-day critical density of $\rho_c = 10^{-29}$ g cm^{-3}. This model is characterized by a single parameter, the time θ which separates the period of anisotropic deformation from the period of practically isotropic expansion. It is evident that if $\theta < 1$ s, then all the important reactions proceed during a stage when the model is already expanding isotropically. The result of the nuclear reactions is then the same as in a Friedmann model. Of course, different results are expected for this anisotropic model if $\theta > 1$ s.

The first computations of nucleosynthesis in anisotropic models were done by Hawking and Taylor (1966) and by Thorne (1967). These computations were carried out under the "canonical assumptions" concerning the particles, namely, that there are no unknown particles, that there is not a large lepton charge, and without an account of nonequilibrium neutrino processes (see § 21.2–4). We shall present an estimate of the amount of

primordial helium produced under these assumptions. For the stage $t \sim \theta$, the expression

$$\tau_{\exp} = \frac{10^3}{T_9^3 \theta^{1/2}} \tag{21.5.11}$$

holds for $\tau_{\exp}$ rather than the expression (21.5.8) where θ is in seconds. This formula is easily obtained from equation (20.1.4) with the substitution of $\epsilon = \rho c^2 = \sigma T^4$ and it is solved with respect to time. Equating $\tau_{\exp}$ from expression (21.5.11) and τ_{weak} from expression (21.5.6), we find the freeze-out temperature to be $T_9 = 10\,\theta^{1/4}$ if $\theta \geq 1$ s. We then find

$$\left(\frac{n_n}{n_p}\right)_{\text{freeze}} \approx \exp\left(-1.5\theta^{-1/4}\right) \tag{21.5.12}$$

for the value of (n_n/n_p) that is frozen in. Subsequently, the neutrons are captured by protons.

Substituting equation (21.5.12) into equation (21.25.10), we find that the helium abundance can be 100% in models with $\theta \gg 1$ s. For very large values $\theta > 10^{11}$ s, however, the expansion during the stage when the neutrons are captured proceeds so rapidly (the time $\tau_{\exp}$ in formula (21.5.11) for a fixed T_9 and very large θ is small) that neutron capture cannot occur. Thus, as Thorne's (1967) calculations show, we have almost a complete absence of ^{4}He in the primordial matter for $\theta > 10^{11}$ s. The dependence of the primordial abundance of ^{4}He on the parameter θ is shown in Figure 56. One can also examine the abundances of the other light elements. Results for them are also displayed in Figure 56.

For realistic calculations involving anisotropic models, however, it is necessary to take account of the processes involving neutrinos considered in this chapter. In this case formula (21.5.11) is valid only before the freeze-out time. After this the entropy grows, so that the real time of

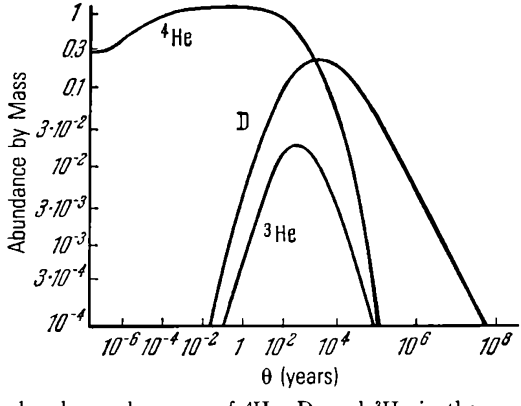


FIG. 56.—The abundances by mass of ^{4}He, D, and ^{3}He in the prestellar matter, as functions of the anisotropy parameter θ.

isotropization, θ^*, does not coincide with θ. Thus, for the parameter value $p_1 = -\frac{1}{3}$ in the Kasner model (see the previous section),

$$\theta^* \approx \theta^{7/16} . \tag{21.5.13}$$

As rough estimates made under the simplest assumptions show, the presence of directed fluxes of energetic neutrinos (affecting the dependence of τ_{exp} on T and θ) does not substantially change the particular processes leading to the synthesis of ^{4}He.

In agreement with the rough estimate, if one pays attention to all of these facts, then the curve for ^{4}He in Figure 56 does not change qualitatively, though its right-hand slope shifts somewhat to the left. (In Fig. 56, θ represents the real time of isotropization.) However, the processes involving the interaction of the fluxes of neutrinos with the synthesizing ^{4}He cannot lead to the formation of a noticeable amount of other light elements. We note, though, that concrete calculations to demonstrate this are, unfortunately, lacking.

Estimates of ^{4}He formation carried out for the case of more complicated anisotropic models, ones with a curved three-dimensional space and with a density less than the critical density, show analogous changes. Here we are not considering the possible influence of hypothetical weakly interacting particles such as gravitons and are limiting ourselves to the simplest assumptions concerning the properties of neutrinos. An account of all of these facts will still not change the conclusions qualitatively. As a whole, the qualitative picture remains as before. As we have already seen in § 7.6, the abundance of primordial ^{4}He is not greater than 30% and is probably close to this value. Thus, the chemical composition of the primordial matter imposes strong limitations on the possible parameters of the simplest anisotropic models—on the isotropization parameter θ, for example.

In all of the calculations we have discussed above, it was assumed that the material is at rest with respect to the reference system of the model. In the general case, however, the material is in motion relative to the reference system (see § 20.6). This leads to the appearance of at least one additional parameter in expressions of the type (21.5.11) and (21.5.12). As an example we shall examine the simplest case. In it, the energy density changes according to the law $\epsilon = \epsilon_0 \tau^{-\beta}$ before the time $t = \theta$; afterwards the solution isotropizes and exhibits Friedmannian behavior. We then have

$$\tau_{exp} = \frac{10^{3\gamma}}{\theta^{(3\gamma/2)-1} T_9^{3\gamma}} , \tag{21.5.14}$$

where $\gamma = 4/(3\beta)$. We further find that

$$(T_9)_{freeze} = 10\theta^{(3\gamma-2)/(10-6\gamma)} ;$$

$$\left(\frac{n_n}{n_p}\right)_{freeze} = \exp\left[-\tfrac{3}{2}\theta^{-(3\gamma-2)/(10-6\gamma)}\right] . \tag{21.5.15}$$

Models with matter at rest correspond to the value $\gamma = 1$; models with any value $0 \leq \gamma < \infty$ are in principle possible. Notice that for $\gamma > 1$, the conditions for the formation of ^{4}He are more favorable than for $\gamma = 1$. For $\gamma \to 5/3$, the freeze-out temperature $T \to \infty$. For $\gamma > 5/3$, there is no period at all when equilibrium between the neutrons and the protons exists, so that the primordial ^{4}He abundance is then determined by the initial conditions alone.

Although any value of γ is possible in principle, the most probable values, as we already noted at the beginning of the section, correspond to $4 \leq \beta < \infty$ and satisfy $0 < \gamma \leq 1/3$. Such values correspond to HA models in which homogeneous matter motions at relativistic speeds arise as a result of the instability of these models with respect to such motions.

For $\gamma = 1/3$, we have

$$\left(\frac{n_n}{n_p}\right)_{\text{freeze}} = \exp\left(-\tfrac{3}{2}\theta^{1/8}\right), \tag{21.5.16}$$

while for $\gamma \to 0$, we have

$$\left(\frac{n_n}{n_p}\right)_{\text{freeze}} = \exp\left(-\tfrac{3}{2}\theta^{1/5}\right). \tag{21.5.17}$$

It is apparent from these formulae that the primordial helium abundance is less than $\sim 10^{-2}$ in anisotropic models with matter motion if

$$\theta > (10^3 - 10^4)\ \text{s}.$$

In anisotropic models, therefore, only three "stable" values of the helium abundance in the prestellar matter are possible: (1) a ^{4}He abundance of $\sim 30\%$ for $\theta < 1$ s; (2) a ^{4}He abundance of almost 100% for models with matter at rest and for which $\theta > 1$ s and $100\ \text{s} < \theta^* < 10^{11}$ s; and (3) a negligible ^{4}He abundance for the same models when $\theta \gg 10^8$ s and also for models when moving matter when $\theta \gg 10^3$ s.

Much was already said in § 7.6 concerning the observations of the primordial chemical composition. One can say with certainty that the ^{4}He abundance in the prestellar matter is less than 30%. Therefore, all models which produce a greater ^{4}He abundance are in conflict with observations.

In this section we have not spoken about the limitations on the anisotropic models from the observations of the isotropy of the RR. We shall discuss this in § 22.8 and § 22.9. Questions concerning models with matter motion are actually quite complicated. Their isotropization in the inhomogeneous case has still not been well studied, for example. Note, however, that if macroscopic regions move, they must subsequently collide.

22 A GENERAL ANALYSIS OF HOMOGENEOUS COSMOLOGICAL MODELS

22.1 The Concept of the Homogeneity of a Cosmological Model

Up to now, anisotropic cosmological models with a Euclidean (flat) three-dimensional space have been examined. Such a space is clearly homogeneous with respect to its geometrical properties. A simple homogeneous distribution of matter and its motion in this space were then examined. However, such simple models by no means exhaust all possible homogeneous (though, generally speaking, anisotropic) models. We shall therefore now pass to a systematic examination of other, more complicated cases.

The concept of spatial homogeneity seems intuitively obvious: what is required is an "identity" in the properties of the model at all points in the three-dimensional space at a fixed instant of time. The geometrical properties of the three-dimensional space (its curvature), its deformation over the course of time, its rotation, and the distribution and motion of the matter in it must be the same at all points.

Clearly, the concept of homogeneity is connected with the choice of reference system. If the properties of homogeneity exist in a certain reference system, then a transition to another system will destroy these properties. The question of whether a model is homogeneous therefore amounts to the question of whether there exists a reference system in which the homogeneity conditions hold.

Let us try to present a mathematical formulation of the homogeneity properties. In mathematics, the concept of the homogeneity of a space is well studied. (See, for example, Eisenhart's 1947 book.) In its application to GTR and to cosmology, this concept has also been studied by many authors. We mention here just some of this work: Taub (1951); Zel'manov (1959*b*); Petrov (1966); Heckmann and Schücking (1962); Schücking and Heckman (1958); Hawking and Ellis (1973); Doroshkevich (1968); Belinsky, Lifshitz, and Khalatnikov (1970); and Grishchuk (1967*a*, *c*).

We shall at first only discuss the geometrical properties of a three-dimensional space that are of importance in an examination of homogeneous models. The geometrical properties of two regions in a three-dimensional space, the neighborhoods of the points A and B let us say, are the same if, for each of these neighborhoods, it is possible to choose respective co-

ordinate grids x and x^* in such a way that the metric tensors are the same. In other words, it must be possible to select these grids in such a way that the functional dependence of $g_{\alpha\beta}$ on x_1, x_2, and x_3 in the neighborhood of the point A is exactly the same as the functional dependence of $g_{\alpha\beta}{}^*$ on $x_1{}^*$, $x_2{}^*$, and $x_3{}^*$ in the neighborhood of the point B.

Suppose that a system of coordinates x is specified over an entire space. Furthermore, suppose that the metric $g_{\alpha\beta}(x)$ near the point A is known. To obtain the same geometrical properties at all other points in the space it must be possible to relabel the coordinates in the neighborhood of each point in a special way. It must be possible to make the transformation (dependent on the coordinates of, for example, the point B: b_1, b_2, b_3)

$$x^* = x^*(x, b) \tag{22.1.1}$$

and the consequent metric translation in such a way that the functions $g_{\alpha\beta}(x)$ at point A and $g_{\alpha\beta}{}^*(x^*)$ at point B depend in the same way on their arguments. One can say this somewhat differently. Given the identity of the functions $g_{\alpha\beta}$ of x and $g_{\alpha\beta}{}^*$ of x^*, the transformation (22.1.1) describes a mapping of the space into itself according to which a point with the coordinates x_1, x_2, and x_3 is shifted into another point, whose coordinates are dependent upon b_1, b_2, and b_3. Though the grid of coordinates is transformed according to the law (22.1.1) and though the origin of the coordinates, for example, then lies at another point, the metric $g_{\alpha\beta}$ remains unchanged. A necessary condition for homogeneity is that it must be possible to correlate every point in the space with any other in this fashion.

If, in addition to the metric of the three-dimensional space, other tensors, vectors, or scalars are defined over the three-dimensional space (the stress-energy tensor of the matter, for example), then they too must possess analogous properties. Upon a transformation of the type (22.1.1), the functional dependence of these quantities on the new coordinates must be the same as on the old coordinates.

In order for the properties of homogeneity on three-dimensional spatial sections to hold in a four-dimensional space, it is necessary that the identity condition on the functional dependence hold under transformation (22.1.1) for quantities describing not only the curvature of the three-dimensional space but also the curvature of the four-dimensional space. Notice that if the quantities characterizing a four-dimensional spacetime are homogeneous over the three-dimensional spatial sections, then the stress-energy tensor is automatically homogeneous over the same sections since the tensor can be expressed, with the help of Einstein's equations, in terms of the four-dimensional spacetime curvature tensor.

Suppose that a three-dimensional space is homogeneous. Then, in the neighborhood of each point, one can translate in any direction by a small vector ξ^α; this guarantees that a given point can be joined with any nearby point. Given such a translation of the space at all other points too, the trans-

lation is unambiguously defined; i.e., the vector ξ^α is defined as a function of the coordinates $\xi^\alpha(x)$. It turns out that ξ^α must satisfy the equation

$$\xi_{\alpha;\beta} + \xi_{\beta;\alpha} = 0 . \qquad (22.1.2)$$

The vector ξ_α is called a Killing vector, and the equation (22.1.2) is called the Killing equation. It is evident that in a homogeneous three-dimensional space one can obtain three independent families of Killing vectors by specifying three, noncoplanar, spatial translations in the neighborhood of some point.

The structure of a homogeneous space or, in other words, the nature of various types of homogeneous spaces, is determined by the structure of the Killing vector fields. Such a classification of the homogeneous three-dimensional spaces according to the structure of the Killing vector fields was worked out by Bianchi (1918). We shall present here this classification as applied to cosmology in a somewhat different form, the form presented, for example, in the works of Ellis and MacCallum (1969) and of Belinsky, Lifshitz, and Khalatnikov (1970).

Mathematically, the spatial homogeneity of a cosmological model as described above is formulated in the following manner: A cosmological model is termed spatially homogeneous if it admits a three-parameter group of motions, i.e., transformations which preserve the functional form of the metric while acting transitively on spacelike three-dimensional hypersurfaces. The transitivity stipulation means that the transformation must permit each point to be joined with every other while the three-parameter dependence mentioned is manifested, for example, by the three parameters b_1, b_2, and b_3 in the transformation (22.1.1). It turns out (see, for example, Petrov 1966; Ellis and MacCallum 1969) that the three-dimensional sections on which homogeneity is realized are geodesically parallel. That is, everywhere the sections are the same distance apart. Therefore, if one draws world lines orthogonal at all points to these sections and chooses this congruence of lines as a reference system, then the metric for such a system can be written in the form

$$ds^2 = dt^2 - g_{\alpha\beta}dx^\alpha dx^\beta . \qquad (22.1.3)$$

In other words, such a reference system is "synchronous" according to the terminology of Lifshitz and Khalatnikov. Three-dimensional, t = const., spaces in this reference system are distinguished as homogeneous spaces. It is well known (see Volume 1) that if $g_{0\alpha} = 0$, then the reference system is nonrotating. Thus, all synchronous systems of reference are nonrotating. This, however, does not mean that the matter in a homogeneous cosmological model is not rotating, since the matter can move relative to a given reference system. Of course, the matter must also be homogeneous in a homogeneous model, but, as we shall see later, motion including rotation is not then excluded.

The homogeneity conditions formulated above are called "group-homogeneity" conditions. One can, however, formulate the homogeneity requirement in a different way. Suppose that the expansion in a cosmological model begins with a singularity. We shall reckon time at each point in the matter as the proper time of this element, beginning at the singularity. We shall now call homogeneous those models in which, to an observer, at each element of the medium at one and the same proper time, the Universe appears identical with respect to all properties. As Grishchuk (private communication) has noticed, however, this homogeneity criterion reduces to the group-homogeneity criterion when Einstein's equations are established. That is, all such models are also homogeneous in the group-homogeneity sense. The converse is, generally speaking, not true since the group-homogeneity criterion does not necessarily presuppose the existence of a singularity. (We shall discuss still another homogeneity criterion later.)

Return now to the group-homogeneity criterion. Consider the homogeneous, three-dimensional spaces of models with the metric (22.1.3). Since they are homogeneous, Killing vector fields satisfying equation (22.1.2) can be defined for them. As is shown in the studies cited above, the existence of Killing vector fields for three-dimensional spaces permits one to write the spatial part of the metric (22.1.3) in the form

$$dl^2 = \gamma_{ab}(t)(e_\alpha{}^a dx^\alpha)(e_\beta{}^b dx^\beta) . \tag{22.1.4}$$

Here the $\gamma_{ab} = \gamma_{ab}(t)$ depend only on t, whereas the $e_\alpha{}^a$ depend only on the spatial coordinates. The $e_\alpha{}^a dx^\alpha$ are three differential forms that do not change upon motion, i.e., they do not change their form under the transformation (22.1.1). The (Latin) superscript in the expression $e_\beta{}^a$ enumerates the differential forms implicit in transformation (22.1.1). The independence of the expressions in parentheses in equation (22.1.4) with respect to transformation (22.1.1) also guarantees that $g_{\mu\nu}$ has the same functional dependence on x that $g_{\mu\nu}{}^*$ has on x^*, to use our earlier description. It furthermore turns out that the invariance of the differential forms $e_\alpha{}^a dx^\alpha$ under the transformation (22.1.1) requires that the relations

$$e_a{}^\alpha e_b{}^\beta(\partial_\beta e_\alpha{}^c - \partial_\alpha e_\beta{}^c) = C^c_{\cdot ab} \tag{22.1.5}$$

be fulfilled. Here the $e_\alpha{}^\beta$ are determined by the relations $e_a{}^\alpha e_\alpha{}^b = \delta_a{}^b$, $e_\alpha{}^a e_a{}^\beta = \delta_\alpha{}^\beta$, while the $C^c_{\cdot ab}$ are constants. From formula (22.1.5) it is clear that

$$C^d_{\cdot ab} = - C^d_{\cdot ba} . \tag{22.1.6}$$

In addition, the identity

$$C^a_{\cdot bd}C^s_{\cdot ae} + C^a_{\cdot de}C^s_{\cdot ab} + C^a_{\cdot rb}C^s_{\cdot ad} = 0 \tag{22.1.7}$$

is satisfied.

For a given three-dimensional space, the choice of the vectors $e_\mu{}^a$ is not

unique, just as in the case for the Killing vectors. A linear combination $A_\nu{}^\mu e_\mu{}^a$ with constant $A_\nu{}^\mu$ evidently satisfies the equations (22.1.5) again, but with different constants $C^a_{\cdot bd}$. Classifying the homogeneous spaces thus amounts to compiling a list of groups of so-called structure constants $C^a_{\cdot bd}$ satisfying equation (22.1.5), but such that the groups do not reduce to one another by means of admissible transformations of the $e_\mu{}^a$.

An analysis shows that such a choice for the "basis" can be presented as in Table 16. The components differing from those presented in Table 16 by the order of the subscripts have the opposite sign, as equation (22.1.6) indicates. The remaining components of the $C^a_{\cdot bd}$ basis are identically equal to zero.

The solution of the Killing equations for the group operators permits one to find the explicit dependence of the metric on the spatial coordinates. One can find the dependence of metrics of all types on the spatial coordinates in, for example, the book by Petrov (1966). We shall not present these results in their entirety here but shall note only some important cases.

The three-dimensional type I models are always flat spaces, though the tensor giving the rates of deformation of the spatial sections is, generally speaking, anisotropic. The metric for this type of model is written in the form

$$ds^2 = dt^2 - a^2(t)dx^2 - b^2(t)dy^2 - c^2(t)dz^2 . \qquad (22.1.8)$$

The group motions are here parallel translations of the three-dimensional space along the coordinate lines of a Cartesian coordinate system. The properties of models of the type (22.1.8) were considered in detail in the preceding chapters.

Models with a flat three-dimensional space are also a particular case of the type VII models, though in these models the deformation is different (see the previous section).

The three-dimensional spaces of type V models are spaces of constant

TABLE 16

THE STRUCTURE CONSTANTS DISTINGUISHING THE VARIOUS HOMOGENEOUS, THREE-DIMENSIONAL SPACES

Bianchi Type	$C^2_{12}=C^3_{13}$	C^1_{23}	C^2_{31}	C^3_{12}
I	0	0	0	0
II	0	1	0	0
VII_0	0	1	1	0
VI_0	0	1	−1	0
IX	0	1	1	1
VIII	0	1	1	−1
V	1	0	0	0
IV	1	0	0	1
VII_h	a^*	0	1	1
III	1	0	1	−1
VI_h	$a\neq 1$	0	1	−1

NOTE.—The parameter a^* takes on all positive values.

negative curvature, but the deformation of them is, generally speaking, anisotropic over the course of time.

Spaces of constant negative curvature are among the type VII_h models as well.

Finally, closed three-dimensional spaces of constant positive curvature are included among the type IX models.

The classification presented above encompasses all the homogeneous spaces that correspond to the definition given earlier. Grishchuk (1970) has remarked that one ought to call spatially homogeneous not only models admitting three-parameter groups of motions acting transitively on three-dimensional spacelike hypersurfaces, but also models admitting larger groups of motion (four- or six-parameter groups; a full, five-parameter group of motions on a three-dimensional hypersurface is impossible). This includes cases where the model does not contain a three-dimensional subgroup on the three-dimensional hypersurfaces and therefore does not fall in the aforementioned classification.

Actually, an increase in the number of group parameters only increases the symmetry of the space (by adding, for example, rotation to the possible motions of which translation is already one) and does not destroy the homogeneity at all.

The analysis of this question showed that such an enlargement of the concept of homogeneity forces one to add only a cosmological model with a metric having the form

$$ds^2 = dt^2 - c^2(t)dz^2 - a^2(t)(d\theta^2 + \sin^2\theta d\phi^2) . \qquad (22.1.9)$$

The three-dimensional space of this model is the topological product of a sphere and a straight line. Models of such a form have been studied by Novikov (1961), Doroshkevich (1965), Zel'dovich (1965*a*), Grischchuk (1967*a*), Ruban (1971), and others. This model admits a four-parameter group of motions that acts transitively on the three-dimensional spatial sections. In the subsequent development, when speaking about the group-homogeneity criterion, we shall understand it to apply in the enlarged sense and shall include model (22.1.9) in any considerations.

22.2 A Differential Criterion for Homogeneity

In the literature, a homogeneity criterion that differs from the group-homogeneity criterion is encountered. It was suggested by Zel'manov (1959*b*) and is called the differential criterion. According to this criterion, a model is considered to be homogeneous if the three-dimensional covariant derivatives of all scalar, vector, and tensor fields with respect to all of the spatial coordinates vanish.

At first glance this criterion seems just as natural as the group-homogeneity criterion. However, the differential criterion imposes more stringent

demands on models than does the group-homogeneity criterion. Actually, this criterion requires more than just that the properties of the Universe be the same at various points of the three-dimensional space at the same time. It requires also that the orientation of vector and tensor fields not change upon a transition from one point of the three-dimensional space to another at a fixed instant of time.

To elucidate the difference between the two criteria, we shall present an example that is a simplification of one from Grishchuk's (1971) work. Suppose that there is a time-independent flux of test particles (particles that do not produce a gravitational field) in a flat three-dimensional space. We specify the velocities of the particles in a Carteseian coordinate system by means of the formulae

$$\begin{aligned} v_x &= a \cos (Az + B) , \\ v_y &= a \sin (Az + B), \\ v_z &= 0. \end{aligned} \tag{22.2.1}$$

In absolute magnitude, the velocity is everywhere the same. The velocity vectors lie in planes $z = \text{const.}$ and, in each such plane, the velocity does not depend on the coordinates x and y. Upon a transition from one section $z = \text{const.}$ to another $z = \text{const.}'$, the velocity vector rotates, changing its orientation. According to the differential criterion, this model is not homogeneous since the derivative of a velocity vector with respect to the coordinate z differs from zero. On the other hand, this model is homogeneous according to the group-homogeneity criterion. One can displace it in the three-dimensional space in such a way that the velocity field transforms into itself and in such a way that one can join each point with every other. One can parallelly shift the model along the x- and y-axes. One can also shift the model along the z-axis if, at the same time, one rotates it around the z-axis on an angle proportional to z. Indeed, such a group of motions is characteristic of Bianchi type VII models. In any case, the difference between the two criteria is clearly evident in this example.

We emphasize that in nonrelativistic mechanics vector or tensor fields in flat space are sometimes said to be homogeneous when the derivatives of the components of a vector or tensor with respect to the Cartesian coordinates equal zero, i.e., when the differential criterion is satisfied. However, we are interested in the identity of properties at all of the points in the space, and not in whether the vectors and the tensors rotate upon a transition from one point to another point of the three-dimensional space. In the example presented above, the properties of all of the points in the space are clearly the same, and one ought to consider the model homogeneous. In the subsequent development, therefore, we shall take homogeneity to mean homogeneity in the group sense if no special stipulations are made.

Concerning models that satisfy the differential criterion, Grishchuk (1971) has shown that all such models must be either isotropic Friedmann models, anisotropic models with a flat space of the form (22.1.8), models of the form (22.1.9) or, finally, models whose space is the product of a pseudosphere and a straight line. The metric for the last is written in the form

$$ds^2 = dt^2 - c^2(t)dz^2 - a^2(t)(d\theta^2 + \sinh^2\theta d\theta^2) . \qquad (22.2.2)$$

But this model is a particular case of the Bianchi type III models. Thus, the class of differentially homogeneous models is already included in the class of models satisfying the group-homogeneity criterion. Therefore, all differentially homogeneous models are also homogeneous in the group sense, if one understands the latter as outlined at the end of the previous section.

In concluding this section, let us return to the example of the vector field (22.1.1) in a flat space. This example is methodologically interesting in that it demonstrates graphically how a different specification of the vector fields in the same space can lead to various Bianchi types. In the example under consideration, the space is a three-dimensional flat space and the vector fields are the velocity fields of test particles. As was already noted, a model with the vector field (22.1.1) is characteristic of a type VII model since a translation along the z-axis requires a rotation of the whole space through an angle proportional to z if homogeneity is to be preserved. If, however, one specifies the velocity field in the three-dimensional space to be everywhere the same and parallel to each other, then the model will be related to the type I model. For the classification scheme, therefore, what is important is not only the geometrical properties of the space but also, generally speaking, the properties of all of the vector and tensor fields under consideration too.

22.3 The Dynamical Properties of Homogeneous Models near a Singularity

We shall now turn to the problem of the time-evolution of homogeneous cosmological models. For homogeneous models, this problem is much simplified by comparison with the situation for the general case since the Einstein equations reduce then to a system of ordinary differential equations. To obtain these equations for each point of a three-dimensional spatial section, one introduces a tetrad of vectors $e_\alpha{}^a$, where the $e_\alpha{}^a$ are the same quantities which are introduced in the expressions (22.1.4), and where the Latin superscript numbers the vector. Then, all three-dimensional vector and tensor fields are projected onto this tetrad. For example,

$$T^{ab} = T^{\mu\nu}e_\mu{}^a e_\nu{}^b , \quad R^{0a} = R^{0\mu}e_\mu{}^a , \quad \mu^a = \mu^\mu e_\mu{}^a . \qquad (22.3.1)$$

The raising and lowering of indices is carried out with the help of the

matrix γ_{ab} or with its inverse $\gamma^{ab}(\gamma^{ab}\gamma_{bc} = \delta_c{}^a)$, so

$$T_{ab} = T^{cd}\gamma_{ac}\gamma_{bd} . \tag{22.3.2}$$

By performing a similar projection of the Einstein equations, we are led to the following system of ordinary differential equations (see Schücking 1963 for a derivation):

$$-\frac{d}{dt}\mathfrak{D}_a{}^a - \mathfrak{D}_a{}^b\mathfrak{D}_b{}^a = \frac{8\pi G}{c^4}(T_0{}^0 - \tfrac{1}{2}T) , \tag{22.3.3}$$

$$-\mathfrak{D}_b{}^c(C^b_{\cdot ca} - \delta_a{}^b C^f_{\cdot fc}) = \frac{8\pi G}{c^4}T_a{}^0 , \tag{22.3.4}$$

$$-\frac{1}{\gamma^{1/2}}\frac{d}{dt}(\gamma^{1/2}\mathfrak{D}_a{}^b) + \gamma^{bc}(A^d_{\cdot af}A^f_{\cdot cd} + C^d_{\cdot df}A^f_{\cdot ac}) = \frac{8\pi G}{c^4}(T_a{}^b - \tfrac{1}{2}\delta_a{}^b T) . \tag{22.3.5}$$

The notation here is

$$\mathfrak{D}_{ab} = \frac{1}{2}\frac{d}{dt}\gamma_{ab} , \tag{22.3.6}$$

$$A^c_{\cdot ab} = \tfrac{1}{2}(C^c_{\cdot ab} + C^f_{\cdot bd}\gamma_{fa}\gamma^{cd} - C^f_{\cdot da}\gamma_{fb}\gamma^{dc}) . \tag{22.3.7}$$

Note that $\gamma \equiv \det \gamma_{ab}$. To this system is added an equation of state that does not violate the homogeneity condition, one possibility being the equation of state for collisionless particles.

All of the quantities in equations (22.3.3–5) depend only on time and do not depend on the spatial coordinates; the C^q_{ab} are constants. Substituting into equations (22.3.3–5) the values of the structure constants from Table 16, we obtain the equations that describe the time evolution of the various types of models.

We shall discuss first of all the behavior of the models near a singularity. Let us therefore pose the following question: Is there an asymptotic solution having the form of equations (19.3.6) and (19.3.7) near the $t \to 0$ singularity and applicable also in the general case of anisotropic models with a curved homogeneous space? In other words, does the solution near the singularity in the general case exhibit expansion along two directions and contraction along the third when expansion from the singularity is being considered? An analysis of this question was carried out first by Lifshitz and Khalatnikov (1963*a*, *b*) and then by Doroshkevich (1968), Grishchuk (1970), Ellis and MacCallum (1969), and MacCallum (1971).

To see what these investigators learned, we now turn to equations (22.3.3–5). Equation (22.3.3) does not contain the structure constants C^q_{ab} which determine the Bianchi type of the model. It moreover does not differ from the equation for models with a flat space, models investigated in detail above. Equation (22.3.4), meanwhile, does not contain second derivatives with respect to time of the sought-for quantities $\gamma_{ab}(t)$. These equations merely impose constraints on the choice of possible initial con-

ditions, and, as is well known (see, for example, Petrov 1966), these equations do not enter directly in the integration with respect to time. If these equations are satisfied, by virtue of the choice of initial conditions, and if the remaining equations are satisfied in a four-dimensional region, then equations (22.3.4) too are automatically satisfied over the same region. Finally, we note that the structure constants C^q_{ab} enter in an important way in equations (22.3.5), which we shall now consider.

Assume that the asymptotic solution near the singularity has the character of equations (19.3.6) and (19.3.7). The terms on the left-hand side of equations (22.3.5) are of order $\sim t^{-2}$. Concerning the stress-energy tensor, therefore, one can repeat what was said in § 19.3: There is, generally speaking, a vacuum stage near the singularity for which one can neglect the right-hand side of equations (21.3.5) as $t \to 0$ since those terms depend on t raised to a power greater than -2.

The fulfillment of the conditions of the form (19.3.6) and (19.3.7) guarantees the fulfillment of the equations (22.3.3). It also guarantees that the first terms in equations (22.3.5) go to zero as $t \to 0$. In the terms in equations (22.3.5) that contain the structure constants C^q_{ab}, there can, generally speaking, appear terms of an order higher than t^{-2}. This is related to the appearance of a scale factor for one direction with a negative exponent for t in the regime determined by equations (19.3.6) and (19.3.7). Suppose, for example, this is the direction along the tetrad with the index 1. Then a term of the form $C^1_{23}(\gamma'_{11})^2\gamma'$ appears, where γ is the determinant of the matrix γ_{ab}. This term is then of order $\sim t^{-2(p_2+p_3-p_1)} = t^{-2(1+2|p_1|)}$. The quantity in parentheses in the exponent is greater than unity, i.e., the whole term is of an order greater than t^{-2}, so equations (22.3.5) are not satisfied. If these equations are to be satisfied, it is necessary that some constant C^q_{ab} with $a \neq b \neq g$ be equal to zero. This corresponds to a tetrad with the index g denoting the axis along which the scale factor has a negative exponent for its t dependence. So, for $t \to 0$, a negative exponent for t along the tetrad direction with the index 1 is possible only if $C^1_{23} = 0$, and so forth.

One glance at Table 16 is sufficient to convince oneself that at least one constant C^q_{ab} with $a \neq b \neq g$ (the last three columns of the table) is zero for the first seven types of models. (Note that the Bianchi types are not listed in order in Table 16.) This makes possible the establishment of asymptotic behavior of the form (19.3.6) and (19.3.7) for these models, behavior with a negative exponent on t in the scale factor for one direction.

Thus, asymptotic behavior similar to what occurs for the Kasner solution is a general feature of the first seven Bianchi types when $T_0{}^\alpha = 0$. If $T_0{}^\alpha \neq 0$, the asymptotic behavior could be different from the Kasner solution (see Lukash 1974*b*; Lukash, Novikov, and Starobinsky 1975) and is analogous to the model of § 22.4.

In the approach to the singularity, but before the establishment of the

asymptotic regime of the Kasner type, the properties of the solution are, of course, more complicated for models with a curved space than for models with a flat space. Recall that for models with a flat space but with a magnetic field, the stress-energy tensor terms in the equations may not have asymptotic behavior with a negative exponent for t characterizing the scale factor along the field direction (see § 20.2). Similarly, terms involving the curvature of the space force one to modify the solution for $t \to 0$ until such time as a Kasner regime with $p_i < 0$ in an "allowed" direction is established. We shall discuss the intermediate stages in more detail in the next section.

Let us immediately take the stipulation that the solution of the GTR equations for the first seven Bianchi types listed in Table 16 also permit one to admit as particular degenerate cases solutions with asymptotic behavior near the singularity differing from that of the Kasner solution. The Friedmann solutions are examples of such degenerate cases. Note, however, that all of these degenerate cases require special initial conditions. We shall not be examining them in detail but shall refer the reader for details to the works cited above. An exception is the Friedmann model; special reasons indeed exist to study this "degenerate" case. It was of course studied in earlier parts of the book.

The type VIII and type IX models do not admit Kasner asymptotic behavior near the singularity since for them all of the structure constants of the form C^g_{ab} with $a \neq b \neq g$ differ from zero. The evolution of models of this type will be studied in the following sections.

22.4 The "Mixmaster" Model

Let us return to one of the questions posed in the introduction to this part of the book. As was remarked there, the homogeneity and isotropy of the Universe over large scales require a special explanation. Suppose, for example, that the homogeneity arises as a result of some kind of physical processes taking place at the beginning of the cosmological expansion. Then, as a minimum, it is required that points separated by a distance corresponding to the scale of the homogeneity be causally connected at the time when the equalizing processes occur. In other words, a signal moving at the speed of light must manage to cover this distance during the period from the singularity to that of the equalizing process. Of course, this condition is by no means sufficient for the smoothing of inhomogeneities. What is more, Zel'dovich has shown (1970*a*) that the smoothing of all of the inhomogeneities does not take place. The requirement mentioned is, however, necessary.

The presence of an optical horizon in a homogeneous, isotropic Friedmann model makes the equalization of the inhomogeneities—the mixing over appreciable scales at the beginning of the expansion—impossible in principle.

But Misner (1969*a*, *b*) proposed a cosmological model in which an optical horizon is absent, so mixing is possible. The Misner model is based on the solution of the Einstein equations for the Bianchi type IX models, a solution discovered by[1] Lifshitz, Khalatnikov, and Belinsky (see Khalatnikov and Lifshitz 1970; Belinsky, Lifshitz, and Khalatnikov 1970; and Belinsky and Khalatnikov 1969*a*, *b*).

Misner's idea was the following. Suppose that a model is described by a Kasner solution with $p_1 = p_2 = 0$ and $p_3 = 1$ for some interval of time. In this case one obtains the equations

$$dt = tdz\,; \quad z = \int_{t_0}^{t} \frac{dt}{t} \sim \ln \frac{t}{t_0} \tag{22.4.1}$$

for the propagation of light along the z-axis ($c = 1$). If $t_0 \to 0$, $z \to \infty$. Thus, light manages to move as great a distance as is desired along the z-axis if t_0 is sufficiently small. Meanwhile, light manages to move only the small distance, proportional to Δt, along the other two axes. If the solution is thereafter rearranged in such a way that $p_1 = 1$ and $p_2 = p_3 = 0$, then the light can move a great distance along the x-axis, and so on. Misner called this the "mixmaster" property.

Misner showed that the Bianchi type IX model, with a closed three-dimensional space, is in principle such a model. It possesses just those properties which allow light to circumnavigate the Universe many times in all directions. In this so-called mixmaster model, even the solution near the singularity is not described by the Kasner model; oscillations along the various axes are superposed on the overall expansion. That the solution for the type IX model cannot be of the Kasner type (true also for the type VIII model) is clear from Table 16 and from equations (22.3.3-5). Actually, all of the C^g_{ab} with $a \neq b \neq g$ differ from zero for these types, so the establishment of Kasner asymptotic behavior in them is impossible (see § 22.3).

The behavior of the solution of this mixmaster model of the Universe near the singularity has been studied in works by Khalatnikov and Lifshitz (1970); Belinsky and Khalatnikov (1969*a*, *b*); Misner (1969*a*, *b*); Belinsky, Lifshitz, and Khalatnikov (1970); Lifshitz, Lifshitz, and Khalatnikov (1970); Doroshkevich and Novikov (1970); S. Novikov (1972); Belinsky, Lifshitz, and Khalatnikov (1971); Grishchuk, Doroshkevich, and Lukash (1971); and Matzner, Shepley, and Warren (1970). The last three studies address the general case in which matter moves relative to the system of reference. However, this condition does not change the basic features of a solution, so we shall consider here only the case in which the material is at rest. Briefly stated, the analysis shows that the solution near the singularity differs from the simplified picture outlined above for other models.

The metric of the mixmaster model is written with a synchronous refer-

1. Note added by Kip S. Thorne. It was also discovered, independently, by Misner (1969*a*).

ence system in the following way:

$$g_{00} = -1\,, \quad g_{\mu\nu} = \gamma'_{ab} e_\mu{}^a e_\nu{}^b\,;$$

$$\gamma'_{ab} = \begin{pmatrix} a^2(t) & 0 & 0 \\ 0 & b^2(t) & 0 \\ 0 & 0 & c^2(t) \end{pmatrix}; \qquad (22.4.2)$$

$$e_\mu{}^a = \begin{pmatrix} \cos z & \sin z \sin x & 0 \\ -\sin z & \cos z \sin x & 0 \\ 0 & \cos x & 1 \end{pmatrix}.$$

The functions a, b, and c describe the anisotropic deformation at each point. Their magnitude characterizes the size of the closed model in three orthogonal directions. The changes in a, b, and c of course describe the time-evolution of the model.

We have already remarked above that one can neglect the stress-energy tensor when describing the vacuum stage near the singularity. Let us then rewrite the Einstein equations for the type IX model during the vacuum stage, when the stress-energy tensor can be neglected. Concerning the late stage of expansion, when T_{ik} becomes substantial, see § 22.7. They become

$$\begin{aligned} &\frac{1}{abc}\frac{d}{dt}\left(bc\frac{da}{dt}\right) + \frac{a^4 - (b^2 - c^2)^2}{2(abc)^2} = 0\,; \\ &\frac{1}{abc}\frac{d}{dt}\left(ac\frac{db}{dt}\right) + \frac{b^4 - (a^2 - c^2)^2}{2(abc)^2} = 0\,; \\ &\frac{1}{abc}\frac{d}{dt}\left(ab\frac{dc}{dt}\right) + \frac{c^4 - (a^2 - b^2)^2}{2(abc)^2} = 0\,; \\ &\frac{1}{a}\frac{d^2a}{dt^2} + \frac{1}{b}\frac{d^2b}{dt^2} + \frac{1}{c}\frac{d^2c}{dt^2} = 0\,. \end{aligned} \qquad (22.4.3)$$

The analysis of the solution of Einstein's equations shows that the functions $a(t)$, $b(t)$, and $c(t)$ behave in the manner described below. Figure 57 depicts this behavior. There $\ln t$ is plotted along the horizontal scale, while the values of the functions $\ln a^2$, $\ln b^2$, and $\ln c^2$ are given by the vertical scale.

We shall be considering the behavior of these functions as the singularity is approached. Two functions (a^2 and b^2, say) oscillate over some interval of t that we shall call a large cycle. Meanwhile, the third function, c^2, falls monotonically as the singularity is approached. Then c^2 begins to grow (we are going to the limit $t \to 0$) at the same time as one of the oscillatory functions, a^2 for example, begins to fall monotonically. Subsequently, b^2 and c^2 oscillate, while a^2 falls monotonically. After some time, there again occurs an exchange between the oscillating and monotonically changing functions, and so on. The behavior is therefore divided into large cycles.

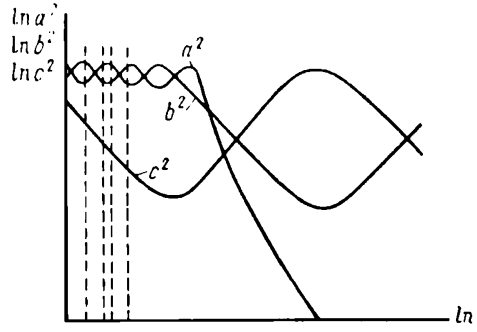


FIG. 57.—The evolution of the "mixmaster" model of the Universe as the singularity is approached.

The volume *abc* meanwhile changes monotonically. Over the small intervals when all three functions change monotonically, the solution is described by the Kasner formulae (see Belinsky, Lifshitz, and Khalatnikov 1970 for more details).

We will use the notation $Q \equiv a^2/b^2$ at the points where the function a^2 is a maximum and $P \equiv a^2/b^2$ at the point where the function c^2 is a minimum in a given cycle. The analysis shows that as the singularity is approached within a large cycle, the amplitude of the oscillations grows. Therefore, if Q is small at the beginning of a given cycle, it becomes large either during the same or (with the exception of the degenerate case which is discussed later) the following cycle. The behavior of the model depends in an important way on still another quantity, on the parameter $\mu = k + x > 1$. We denote by k the integral part $[\mu]$ of the parameter μ. It equals the number of oscillations in the given cycle. The fractional part of μ, x, determines the parameter μ for the next cycle by $\mu_{s+1} = x_s^{-1}$, where s is the cycle number. Now suppose that $Q \gg 1$. Then

$$P_1 = |4Q_{01}|^{\mu^2}, \tag{22.4.4}$$

where Q_{01} is the oscillation amplitude at the beginning of a certain cycle. (Remember that we are going from large to small values of t.) Upon the passage to the next large cycle, the relation

$$Q_{02} \approx P_1 \tag{22.4.5}$$

is valid, i.e., the oscillation amplitude at the start of the next cycle is of the order of P_1 and is large. For the quantity P_2, we have

$$P_2 \approx P_1^{\mu_2^2}. \tag{22.4.6}$$

Thus the quantity P grows catastrophically. For the large cycle numbered s, we have

$$P_s \approx (4Q_{01})^{\mu_1^2\mu_2^2\cdots\mu_s^2}. \tag{22.4.7}$$

We emphasize that this formula in practice describes the extent of the entire change of the function, which, during cycle s, is the smallest function. It describes this change for the duration of all of the cycles from cycle 1 to cycle s because practically all of the changes in the amplitude of P take place during cycle s, which is the one nearest to the singularity of all of those under consideration [although cycle $(s + 1)$ of course follows cycle s, and so on]. (The number of cycles is infinite if no limitations are imposed on the applicability of GTR. For more concerning this point, see the next section.)

The formulae presented describe the basic characteristics of the solution as the singularity is approached. Upon motion in the opposite direction, the terms in the gravitational equations describing the matter, the terms that we discarded, play an ever growing role. We shall consider in § 22.6 the behavior of the model during the stage when these terms become important.

Let us return to the anisotropic stage. As the singularity is approached, a tendency is observed for growth in the relative amplitudes of the oscillations of the functions (growth in P and Q). This tendency can be stopped only in the following situation. In each cycle, not exactly a whole number of oscillations can be covered, so the numbers $\mu = k + x$ are not integral. A "dangerous" situation can arise if μ is very close to an integral value. This happens if x is small or, in particular, if x or $1 + x$ falls between the values

$$1 \pm (4Q^*)^{-1}, \tag{22.4.8}$$

where Q^* is the oscillation amplitude at the end of a large cycle. If, in addition to this, all of the functions are of the same order at some instant t at the end of the cycle, $a \approx b \approx c$, then oscillations with a small value of Q can arise.

Before proceeding further, notice that what Misner requires for the realization of the idea of a mixmaster model is just the systematic occurrence of such cycles, with a small value of Q and large value of μ, as the singularity is approached. For when Q is also small during the course of a cycle, the properties of the solution are similar to those of the Kasner solution. Under such circumstances light manages to circumnavigate the Universe many times along the axis characterized by the scale factor c, thereby realizing the mixmaster action.

If similar situations are systematically repeated, then the mixing takes place in all directions. It is necessary to emphasize, however, that the systematic occurrence of cycles with small Q is quite improbable. The probability that, in a given cycle k, the integral part of μ lies between the values $[1 \pm (4Q^*)^{-1}]$ is

$$W = (4Q^*)^{-1} \tag{22.4.9}$$

This last quantity decreases tremendously rapidly from cycle to cycle as the singularity is approached. The total probability for encountering a dangerous zone (even if this happens once!) in some cycle right up to the singularity is determined by the sum of a series:

$$W \approx (4Q_1^*)^{-1} + (4Q_2^*)^{-1} + \ldots . \qquad (22.4.10)$$

The sum of this series is essentially equal to its first term, and, since $(4Q_1^*)^{-1} \ll 1$, the sum is small.

The probability for a systematic repetition of the Misner situation is thus vanishingly small. Note too that we are here not discussing the question of whether or not one can apply the mixmaster model of the Universe right up to the singularity without limitation. The problem there is that when the spacetime curvature achieves the value $(10^{-33}\,\text{cm})^{-2}$, quantum gravitational effects begin to play a role. We shall make several remarks in this connection at the end of the next section. Here, however, we emphasize simply that even without a limitation on the applicability of GTR, the value of Q always grows as the singularity is approached, given a random choice for the parameters of the model.

In concluding this section, notice that the Bianchi type VIII model possesses evolutionary properties near the singularity similar to those just discussed. But this model possesses instead a three-dimensional space with the topology of a three-dimensional cylinder.

22.5 On the Impossibility of Mixing in the Mixmaster Model

We shall now turn to the question of whether a signal moving at the speed of light can circumnavigate a Bianchi type IX model (Doroshkevich, Lukash, and Novikov 1971; Grishchuk, Doroshkevich, and Lukash 1971).

First of all, the null-geodesic equations show that light rays leaving from some point in the directions of the principal axes of the three-space curvature tensor always move along in these directions subsequently. To realize Misner's idea, it is necessary that such light rays manage to circumnavigate the Universe a large number of times in all directions. Now the changes in the scales of the model along the principal axes of the curvature tensor are described by the functions $a(t)$, $b(t)$, and $c(t)$, respectively. Thus, knowing these functions, one can calculate how many times light can circumnavigate the model in any given direction after a definite interval of time.

If, during a certain cycle, the amplitude Q is small, then light can circumnavigate the model many times in the axis direction characterized by the smallest scale factor. We have seen, however, that with a huge probability, the small amplitudes for Q never arise near the singularity. The question then arises of whether mixing is possible when $Q \gg 1$.

Consider, therefore, the propagation of light along the three principal

directions of the model. During each cycle, the monotonically changing function is always much smaller than the oscillating functions, and the most favorable conditions for circumnavigation by light exist for a ray moving along the axis characterized by this monotonic function. But the number of circumnavigations of the Universe N_s over the time of a large cycle, along this axis, say c, is determined by the formula

$$N_{s,c} \approx \frac{\mu_s}{2\pi {Q_{0s}}^{1/2}}, \qquad \mu_s \gg 1 . \tag{22.5.1}$$

For the number of concurrent circumnavigations by other light rays along the axes, we have

$$N_{s;a,b} < {Q_{0s}}^{-1/2} .$$

Thus, regardless of the path the light takes, the number of times N_s it circumnavigates the Universe in the course of cycle s satisfies the inequality

$$N_s < \frac{\mu_s}{2\pi {Q_{0s}}^{1/2}} \equiv N_s^* . \tag{22.5.2}$$

Consequently, for the total number of circumnavigations of the Universe N by light over the interval of time from the singularity to cycle $s = 1$, the inequality

$$N < \sum_{s=1}^{\infty} N_s^* \tag{22.5.3}$$

holds.

Suppose now that the model parameters are chosen in a random manner. Under this condition, Doroshkevich, Lukash, and Novikov (1971) show that for large values Q_{01}, the sum

$$\sum_{s=1}^{\infty} N_s^*$$

is small by comparison to unity with a probability that differs negligibly from unity. In addition, they show the following: the probability that this sum satisfies

$$\sum_{s=1}^{\infty} N_s^* < x \ll 1 \tag{22.5.4}$$

differs from unity by an amount W, which itself satisfies

$$W < \frac{1}{500 \ln Q_{01}} + \frac{1}{x {Q_{01}}^{1/2}} . \tag{22.5.5}$$

In the paper of Grishchuk, Doroshkevich, and Lukash (1971), it is shown that these same estimates are also valid in more complicated cases where the matter is in motion.

Thus the expressions presented demonstrate that without an extremely special choice of the model parameters, light cannot circumnavigate the

Universe even once in any direction over the time interval beginning with the singularity and lasting to the end of the anisotropic stage, when Q is large. During the isotropization period of the model, when the self-gravitation of the matter becomes important, light may possibly circumnavigate the Universe once, but this is clearly insufficient for a realization of Misner's mixmaster idea.

22.6 Quantum Limitations on the Mixmaster Model

We have already noted repeatedly that GTR is inapplicable to scales smaller than $l_{\text{Pl}} \sim 10^{-33}$ cm, for at such scales quantum fluctuations of the metric are important. In particular, if some scale of a closed homogeneous model is smaller than l_{Pl} then a description of the behavior of the model on this scale by GTR is incorrect. In a study by Misner (1969*b*), however, the assertion is made that this is not a limitation in the case of the mixmaster model, and that GTR is applicable to scales as small as is desired—even to those characteristic of a true singularity. Misner draws this conclusion from an analysis of the quantization of the expansion of the Universe as a whole. (For details about this, see the next part of the book.) Here we shall only outline what he does. Consider times when the scale of the Universe is much larger than l_{Pl}. Misner demonstrates that since these states correspond to a large quantum number n, which is described by classical, nonquantum equations; the number n is an adiabatic invariant with respect to the approach to the singularity (for time translation into the past). Consequently, he asserts that the role of quantum effects does not grow with a decrease in scale, so there is no limitation whatsoever on the applicability of GTR.

Notwithstanding Misner's remarks, it seems to us that l_{Pl} is all the same the limit of applicability on the nonquantum GTR equations. The point is that on scales $l < l_{\text{Pl}} \approx 10^{-33}$ cm, quantum fluctuations of the metric exist regardless of the state of expansion of the Universe as a whole. For example, there are such fluctuations in the Universe today on scales $l < 10^{-33}$ cm or for times $t < 10^{-43}$ s, independent of the scales of the Universe and independent of the state of its expansion. And, over these small scales, nonquantum GTR is inapplicable today. Indeed, one can repeat this assertion for any moment in the past. Therefore, when any scale in the Universe in the past is less than $l_{\text{Pl}} \approx 10^{-33}$ cm, or when the time elapsed since the singularity is less than 10^{-43} s, the equations of GTR are inapplicable to a description of the evolution on this scale. But these questions are analyzed in more detail in the next part of the book, so we shall not discuss them further here.

Consider the question of the applicability of the mixmaster model to the early stages of expansion of the real Universe. In the general case, when light does not circumnavigate the Universe, the limitations related to the

temporal duration of the evolution are more severe than the limitations on spatial scales. The limitations on the spatial scales are, however, still very stringent, and we shall now discuss them. The characteristic scale of the Universe today is $L_0 \sim c/H_0 \approx 10^{28}$ cm. But, beginning at least with the time in the past when this scale is $\sim 10^{25}$ cm, the observable part of the Universe expands isotropically. As was stated above, on the other hand, the smallest scale for which the nonquantum theory is applicable is $\sim 10^{-33}$ cm or 10^{-43} s. Thus, over the course of the time interval when the Universe is perhaps described by the mixmaster model, its scale changes by not more than $10^{25}/10^{-33} = 10^{58}$ times. Given this result, let us see how many cycles of oscillation are then possible and how many times light can circumnavigate the Universe in various directions. To estimate the possible number of oscillations, we must make use of formula (22.4.6). In order to obtain on the left-hand side a quantity $P^{1/2} = a/c$ less than 10^{58}, given any Q_0 noticeably greater than unity, it is certainly necessary that there be not more than two or three cycles with two or three oscillations in each cycle. If, however, the number of oscillations in a cycle is eight or 10, there cannot be more than one cycle.

Quantum effects in GTR must therefore strongly limit the possible number of oscillations in the neighborhood of the singularity in the Misner cosmological model.

In conclusion, we emphasize once again that repeated circumnavigations of the Universe by light are impossible in the mixmaster model. As is shown above, this is so even without any limitations on the applicability of GTR, i.e., even given a formal continuation of the GTR solution up to the singularity itself.

22.7 The Isotropization of Homogeneous Cosmological Models in the Course of Expansion

In this part of the book we are examining HA models with the goal of possibly applying them to a description of the early stages of the cosmological expansion. Naturally, only models that isotropize, that approach a Friedmann model, in the course of time can pretend to this role. But what should one understand by the terminology "isotropize" or "approach a Friedmann model"?

The most "natural" answer is the following. By isotropization at some time one means the nearness of all of the model parameters—both the geometrical and dynamical parameters and those concerning the distribution and motion of the matter and radiation—to the corresponding quantities in a Friedmann model. Such an approach would wholly satisfy the hardened theoretical cosmologist. It would, however, apparently be too strict if we want to apply our concepts to the real Universe. Why? Observations at present are still very far from the possibility of testing all of the

properties of various models. The astronomer's fundamental weapon for testing degree of isotropy of the expansion is observations of RR and, to a much smaller degree, observations of galaxies and counts of radio sources in various directions in the sky. The last two methods, which we have discussed in earlier parts of the book, involve excessively large errors. Another tool involves an indirect criterion, the chemical composition of the primordial matter. This is useful as well, for, as we saw in § 21.5, the synthesis of the elements is very sensitive to the expansion rate during the period of nucleosynthesis. Of course, agreement between the observed chemical composition and the predictions of the Friedmann model signifies only that the time rate of change of the volume occupied by matter and the rate of expansion of the Friedmann model coincide.

In attempting to solve the problem of isotropization, one ought to keep in mind that, according to astronomical observations, if the Universe is anisotropic from the outset, then the isotropization occurs rather early, for $z \gg 10^3$, and probably for a much earlier time, for which $z \gg 10^9$. The first figure follows from RR observations (see below), while the second follows from the data concerning the chemical composition (see § 21.5).

Thus, the minimum requirements for the homogeneous model to be similar to the Friedmann model with respect to the very important observational properties discussed now are the following. The model must, sufficiently early (before the start of element synthesis), and with a specified accuracy, begin to expand isotropically. This isotropic expansion must then proceed at a pace that is the same as for a Friedmann model. Finally, this expansion property must be preserved throughout the extended time during which the model changes its linear size—the time from the start of the expansion to the present.

We emphasize too that isotropization must take place at a stage when the expansion proceeds at practically a parabolic ($\rho \sim t^{-2}$) rate, so that the density is practically equal to the critical density.

What about the isotropization of such parameters as the three-space curvature in various directions? Evidently we have not up to now mentioned these in our definition of isotropization. We shall simply say that when an analysis of observational data shows that they are substantial, then it will be necessary to introduce them into the definition as well. The definition will then characterize "effective, observational isotropization."

In the literature concerned with the mathematical analysis of anisotropic models are other definitions of isotropization. S. P. Novikov (1972) and Collins and Hawking (1973*a*), for example, define isotropization differently. One must keep this in mind in order to avoid misunderstanding when reading the literature.

Among homogeneous models there exist ones whose properties are never similar to those of a Friedmann model in the course of expansion. There also exist models whose properties approach those of the Friedmann

model at some stage of expansion but for which the expansion again becomes highly anisotropic. Finally, there exist models which, after isotropization, approach the Friedmann model more and more (or, in any event do not diverge from it) as $t \to \infty$.

It seems almost evident that anisotropic models that approach a Friedmann model as $t \to \infty$ can only be included among those Bianchi types that contain the Friedmann model as a particular case. Indeed, an analysis of this point (Grishchuk 1967*a*; Doroshkevich 1968; MacCallum 1971) shows that this is actually so.

Open or flat Friedmann models exist among the Bianchi type I, V, and VII models. Only for models of these types, therefore, is an infinitely close approach to a Friedmann model in principle possible. For other types of models, however, nearness to Friedmann behavior over an extended time interval and with a specified degree of accuracy is possible. It is, of course, necessary to analyze all models that are similar to the Friedmann model for an extended period and to compare them with observations.

A most important problem is to construct a solution for anisotropic models that is valid over the entire interval of time of interest. That is, it is important to trace the whole process from the singularity to isotropization and to connect the initial data applicable near the singularity with the parameters describing the late stage of expansion. This would permit one to explain which of the models with given parameters near the singularity agree with the data of astronomical observations on the degree of isotropy of the cosmological expansion. One would thus solve the question of just how general is the class of initial conditions for anisotropic models that can lead to the presently observable picture of the Universe.

The process of isotropization in type I models was examined in detail in previous sections. We shall therefore begin with an analysis of the isotropization of the type V models. Thereafter we shall study the isotropization in the type VII and IX models. These last three models are, in a well-known sense, the most general ones. The analysis of the evolution of models of other classes, with particular reference to their possible isotropization, is a special case of the analysis for the type VII and IX models. We shall thus not dwell on an analysis of the other types. The reader can find a discussion of these questions in the analysis by Doroshkevich, Lukash, and Novikov (1973), by Lukash (1973), and by MacCallum (1971). Let us therefore begin our discussion with the type V models.

The three-dimensional space of the type V model is always a space of constant negative curvature. Consider at the outset models with the stress-energy tensor of an ideal gas and with the material at rest in the homogeneous three-dimensional space of the model. We are thus supposing that $T_0^\alpha \equiv 0$. Then, as was shown by Doroshkevich (1968), the solution always reduces to the Heckmann-Schücking (1962) solution. The metric for this solution can be written in the form

$$ds^2 = dt^2 - R^2(t)\{dx^2 + e^{2x}[S^2(t)dy^2 + S^{-2}(t)dz^2]\} . \quad (22.7.1)$$

Near the singularity, the solution takes the form

$$R \sim t^{1/3} , \quad S \sim t^{1/\sqrt{3}} \quad (22.7.2)$$

for $p < \epsilon$, i.e., it assumes the Kasner form. The exponents giving the dependence of the scale factors on time near the singularity satisfy $p_1 + p_2 + p_3 = p_1^2 + p_2^2 + p_3^2 = 1$. The smallest exponent, that for the z-axis, for example, is $\frac{1}{3} - 1/\sqrt{3} = -0.24$. For the late stages of expansion, the model is isotropized and has the asymptotic form

$$R(t) \sim t , \quad S(t) = (1 - t_\phi^2/t^2) . \quad (22.7.3)$$

In the case of dust, $p = 0$, the solution can be written in its entirety as

$$\left(\frac{dR}{dt}\right)^2 = 1 + \frac{2GM}{R} + \frac{a_1^2}{3R^4} , \quad S = \exp\left(a_1 \int \frac{dt}{R^3}\right) ,$$
$$\rho = M/(\tfrac{4}{3}\pi R^3) ; \quad M = \text{const.} , \quad a_1 = \text{const.} \quad (22.7.4)$$

The constant t_ϕ^2 in the asymptotic form (22.7.3) is expressed in terms of a_1 as $t_\phi^2 = a_1/2$ in the case of solution (22.7.4).

The asymptotic form (22.7.3) describes the behavior assumed by the model, behavior characteristic of the Milne model (see Part I), when one can neglect the self-gravitation of the matter and when all of the matter expands at a constant rate, with the motion governed by inertia. The scale factor R is then proportional to the time. However, isotropization of the model is possible at a much earlier stage when one cannot neglect the self-gravitation of the matter. In this case the solution at first approaches the Friedmann solution for the open model at a stage when the gravitation of the material in it is still substantial (see § 1.4 and Fig. 2) and thereafter, as for any open solution, transforms into a Milne model. Evidently, what must interest us is just this case, for if the early stage of the expansion in the real Universe is also anisotropic, then the isotropization takes place rather early, when the gravitation of the matter is important.

It is easy to follow the properties we have discussed in solution (22.7.4). In the expression for $(dR/dt)^2$ the first term describes the kinetic energy of the expanding matter, the second describes the matter's gravitational energy, and the third the energy of the free gravitational field. Near the singularity ($R \to 0$), the second and third terms are small compared to the third, and the solution is anisotropic or Kasner-like. If $G^2M^2 \gg a_1$, then, the third term becomes equal to the second term and both are much larger than unity, and the model is isotropized and exhibits an intermediate Friedmann asymptotic form with $R \sim t^{2/3}$. At this time the self-gravitation of the matter is substantial. Perfectly analogous is the situation for $p = \epsilon/3$, the difference being that $R \sim t^{1/2}$ during this regime. Only later does the model exhibit the Milne asymptotic behavior (22.73). If, however,

$G^2M^2 \ll a_1$, then the isotropization of the model occurs when the third term is comparable with unity. The self-gravitation of the matter is then never important, and the model transforms directly into a Milne model from the Kasner stage.

As we see, the type V models with $T_0^\alpha = 0$ only admit the single asymptotic solution (22.7.2) near the $t \to 0$ singularity. This limitation is removed if one admits the possibility that the matter moves relative to the reference system. At the end of § 20.5 we have already remarked that if one assumes that at some instant of time during the vacuum stage all components of the velocity are of the same order of magnitude, then, as $t \to 0$, the largest velocity component is the one along the axis for which the scale factor has the largest positive exponent in its time dependence. The matter moves in this direction at a relativistic velocity as $t \to 0$. In this case, the law for the change in the density, $\epsilon \sim t$, with respect to the time t in the homogeneous system of reference is

$$\epsilon = \tau^{-4(1-p_1)/(1+3p_1)} \tag{22.7.5}$$

with respect to the proper time τ. However, the type V model still imposes strong limitations on the parameters describing the expansion and the motion of matter in it. Thus, the specification of the parameters of the solution for this model at some time uniquely determines the matter flux in the model. This feature is related to the isotropic character of the curvature of the three-dimensional space of the model, which requires an exact balancing of the separate factors influencing the curvature.

Keeping in mind what has been stated, we shall not dwell any further on the type V model, but shall instead pass to the more general type VII and IX models, which possess anisotropic curvature. We shall present a summary of the analysis by Doroshkevich, Lukash, and Novikov (1973).

We begin with the type IX model. We assume that the material is at rest in the reference system and that the equation of state is $P = \epsilon/3$. A picture of the isotropization appears in Figure 58. As we saw in § 22.4, the evolution of a model of this type near the singularity consists of oscillations of the axes on the background of overall expansion in proportion to the distance away from the singularity! At some instant t^*, which is an arbitrary parameter of the model, the terms describing the gravity of the matter begin to influence the solution. In Figure 58 is depicted the logarithm of the scale factors along the various axes. As is apparent in Figure 58, the two scale factors are comparable after the intermediate stage. They then begin to oscillate with a small and rapidly decaying amplitude, with the decay governed by a power law. After this, all of the axes expand according to a Friedmann dynamical law—accurate up to logarithmic factors. Subsequently, over the course of an extended period of time, and practically up to the time of greatest expansion when the expansion changes into contraction, the expansion is "almost" isotropic. The stipulation

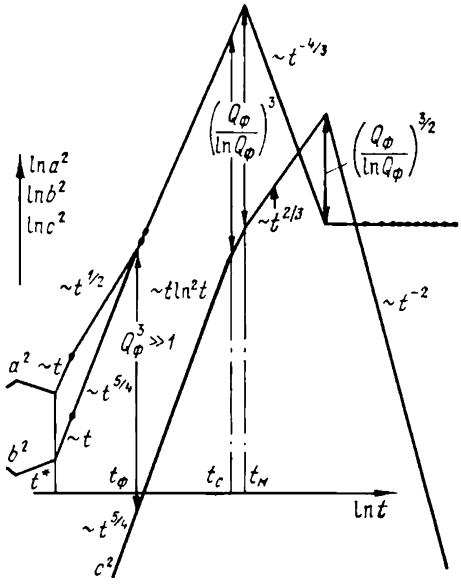


FIG. 58.—The process of "isotropization" in a Bianchi type IX model. This figure shows only the evolution during one change from expansion to contraction.

"almost" means after a moment t_ϕ, when $a \approx b$, the anisotropy of the rate-of-deformation tensor is small and decays, though very slowly, in proportion to $[\ln(t/t_\phi)]^{-1}$. (Concerning the effect of this on the RR anisotropy, see § 22.9.) We want to stress that the curvature of the three-dimensional space is very anisotropic and, as a rule, never becomes isotropic in the future. Only after the equation of state changes, when it effectively becomes $P = 0$ (denote the moment by t_c), do all deviations from exact isotropy decay via a power-law. Finally, the terms describing the average curvature of the three-dimensional space begin to affect the solution (denote this moment by $t = t_M$). The expansion again becomes anisotropic, the volume of the model passes through a maximum, and the model then contracts to a second singularity. Of most interest is that, for arbitrary initial conditions, the gravity of the matter does not automatically lead to an isotropization of the solution wherein all of the quantities a, b, and c are practically equal to one another over the course of an extended period of time. For such a form of isotropization, a quite special selection of initial conditions is necessary.

Given arbitrary (over a wide range) initial conditions, there is an extended period in the solution when the expansion proceeds almost isotropically (with small corrections). The solution is then nearly the same as the critical-density Friedmann solution. However, the principal values of the curvature tensor remain different, and $a \approx b \gg c$.

Consider now the type VII model. The evolution of models of this type

differ from that described above at the two extremes of the time axis. Near the singularity, the solution for the type VII model is Kasner like (this is for $T_0^a = 0$; for the more complicated case see Lukash 1974*b*); then oscillations arise, though there is only one large cycle of the nature outlined in § 22.4; the isotropization described above for the type IX model there upon takes place. For a long time the model expands according to Friedmann modes. But the curvature of the three-dimensional space is very anisotropic. Finally, during the last stage of evolution, when the average curvature of the model begins to influence the solution, the type VII solution differs from that described above.

Note first of all that it is possible here that the average curvature equals zero and, as $t \to \infty$, the model approaches the critical-density, flat Friedmann model. But when the density is less than the critical density, a last stage ensues. The density cannot be greater than the critical density since the model would then be a type IX model.

We now pass to a description of the last stage of evolution. During this stage, the model is close to a Milne model. In this model the amplitudes of the anisotropy of the three-dimensional space and of the anisotropy of the Hubble parameters in the different directions are "frozen." This conclusion has been reached by Collins and Hawking (1973*a*); see also Doroshkevich, Lukash, and Novikov (1973):

$$\Delta H_i/H = \text{const.} \tag{22.7.6}$$

An order-of-magnitude estimate is

$$\frac{\Delta H_i}{H} \approx \left(\frac{t_c}{t_M}\right)^{2/3} \frac{8}{8 + \ln (t_c/t_\phi)}. \tag{22.7.7}$$

It is important to note that solution (22.7.6) does not tend to a Friedmann solution but rather represents a perturbation of the Friedmann solution frozen-in during the Milne stage. Actually the perturbations of the curvature as well as the relative difference in the Hubble "constants" are constant here. True, the ratio c/a grows without bound as $t \to \infty$. However, as we already remarked, the question for the type VII model admits a scale transformation, so, by itself, the growth in this ratio still does not signify anything. More precisely, the model is at any instant of time t described by a Friedmann solution plus small corrections. It will be shown in § 22.4 that the RR anisotropy during this stage is small. Notice that to observe the effect of unlimited growth of the ratio c/a as $t \to \infty$ is in principle possible. In fact, a sphere of particles at rest or moving slowly in the synchronous reference system must in the course of time transform into an ellipsoid with indefinitely growing elongation. In practice, however, it is impossible to conduct such an experiment.

In conclusion, let us make still another remark about the early stages of the evolution of the type VII model. We shall remove the constraint that

the matter be at rest in the system of reference. As was shown in § 20.5, the matter becomes relativistic during the Kasner stage of evolution due to instability. Then, after the vacuum stage and the oscillations resulting from the anisotropy of the spatial curvature, a stage ensues wherein the term describing the anisotropic flux of matter relative to the reference system of the model affects the solution. As one can show, the solution is described by the formulae $a \sim t$ and $b \sim c \sim (\ln t)^{1/2}$ during this period. The flux is oriented along the axis that is contracting. Moreover, the gravitational effect of this flux is such that the model begins to expand along this very axis (the axis characterized by a) and is very slowly changing along the other two axes. This leads to a rapid fall of the velocity and to the assumption by the solution of behavior characteristic of stages mentioned above.

It is in principle possible to select the initial conditions in such a way that the separate stages disappear or are otherwise combined. We have merely described the most general cases of time evolution of the type IX and VII models. As was remarked at the beginning of this section, the model parameters were chosen in such a way that the models isotropize rather early.

The conclusion to be drawn from this analysis is that for a wide range in the choice of parameters of the models (with respect to generality, coinciding with a complete selection of the arbitrary parameters determining a model), a given model isotropizes according to the character of the deformation and, in this sense, approximates a Friedmann model. For another approach with a different definition of "isotropization" and with different conclusions, see S. Novikov (1972), Bogoyavlensky and S. Novikov (1973), and Collins and Hawking (1973*a*, *b*). Note, finally, that the observational consequences of the effects analyzed here are discussed in § 22.9. For other aspects of the evolution of these models see Grishchuk, Doroshkevich, and Udin (1972), Berger (1972), Grishchuk (1973) and Lukash (1974*a*, *c*). Lukash (1974*b*) demonstrated that type VII models can be represented as circularly polarized gravitational waves in nonstationary three-dimensional spaces.

22.8 The Anisotropy of the Relic Radiation (RR) in Bianchi Type I Models with the Critical Matter Density

In this section the question of the RR anisotropy is examined for the very simple type I model which has a flat, comoving three-dimensional space. Analyzing this simple case will permit us to isolate a number of important features of the problem (Doroshkevich, Zel'dovich, and Novikov 1967*c*). In the following section more complicated models will be considered.

Observations show that the large-scale RR anisotropy $\Delta T/T$ satisfies

$$\Delta T/T \leq 3 \times 10^{-3}\,.$$

The Universe thus becomes transparent to RR at a stage when the expansion anisotropy is small. Consequently, for a comparison of the theory with observations, it is necessary to deduce formulae for the expansion of the model to describe the period when the model differs little from a Friedmann model.

We shall write formulae describing the process of isotropization of the model, formulae that take into account the possible presence of an ordered magnetic field as well as of oriented fluxes of relativistic particles. We should emphasize that the presence of a magnetic field in an anisotropic model is not necessary (its presence depends on the choice of initial conditions). But, as was shown in § 21.1, anisotropic fluxes of weakly interacting particles inevitably develop as a result of processes during the early stages of anisotropic expansion. An account of such fluxes is, therefore, definitely necessary.

Suppose that the magnetic field is oriented along the z-axis and that W is its energy density. Suppose further that p_x, p_y, and p_z are the pressures of the free particles along the x-, y-, and z-axes and that ϵ and P are the energy density and pressure of the ordinary matter. Finally, assume that $\epsilon \gg W$, that $\epsilon \gg p_x$, p_y, p_z, and that $a^{-1}da/dt \approx b^{-1}db/dt \approx c^{-1}dc/dt$. The last set of relations of course means that to a first approximation the expansion takes place isotropically. It then follows from equations (19.3.2–5) and from § 20.3 and § 21.1 that

$$\frac{1}{abc}\frac{d}{dt}\left[\left(a^{-1}\frac{da}{dt} - b^{-1}\frac{db}{dt}\right)abc\right] = \frac{8\pi G}{c^2}(p_x - p_y), \qquad (22.8.1)$$

$$\frac{1}{abc}\frac{d}{dt}\left[\left(a^{-1}\frac{da}{dt} - c^{-1}\frac{dc}{dt}\right)abc\right] = \frac{8\pi G}{c^2}[(p_x - p_z) + 2W]. \qquad (22.8.2)$$

We shall presume that $P = \epsilon/3$, whence $a \approx b \approx c \sim t^{1/2}$. In this case,

$$\begin{aligned} \frac{a^{-1}da/dt - b^{-1}db/dt}{a^{-1}da/dt} &= 3\frac{p_x - p_y}{\epsilon} + 2\left(\frac{t_1}{t}\right)^{1/2} \\ \frac{a^{-1}da/dt - c^{-1}dc/dt}{a^{-1}da/dt} &= 3\frac{p_x - p_z}{\epsilon} + 6\frac{W}{\epsilon} + 2\left(\frac{t_1}{t}\right)^{1/2}. \end{aligned} \qquad (22.8.3)$$

Here t_1 is a parameter of the model.

It follows from the formulae that in the absence of a magnetic field and of a flux of relativistic particles the deformation anisotropy decays as $t^{-1/2}$. In the general case, the terms connected with the magnetic field and the anisotropic flux are the fundamental causes of the deformation anisotropy. For $P = \epsilon/3$, the ratio $(p_x - p_y)/\epsilon$ or W/ϵ remains constant at the stage when the expansion is almost isotropic. Therefore, an important conclusion is that in the presence of a magnetic field or a flux of relativistic particles the deformation anisotropy is conserved during the stage when

$P = \epsilon/3$ it does not decrease. But when the tensor T_α^β is anisotropy-free, the deformation anisotropy falls as $\sim t^{-1/2}$.

For the $P = 0$ case, $a \approx b \approx c \sim t^{2/3}$, and we have

$$\frac{a^{-1}da/dt - b^{-1}db/dt}{a^{-1}da/dt} = 6\frac{p_x - p_y}{\epsilon} + \frac{3}{2}\frac{t_1}{t},$$
$$\frac{a^{-1}da/dt - c^{-1}dc/dt}{a^{-1}da/dt} = 6\frac{p_x - p_z}{\epsilon} + 12\frac{W}{\epsilon} + \frac{3}{2}\frac{t_1}{t}; \tag{22.8.4}$$

$$\frac{a-b}{a} = -6\frac{p_x - p_y}{\epsilon} - \frac{t_1}{t},$$
$$\frac{a-c}{a} = -6\frac{p_x - p_z}{\epsilon} - 12\frac{W}{\epsilon} - \frac{t_1}{t}. \tag{22.8.5}$$

It is apparent from these formulae that the anisotropy related to the anisotropy of T_α^β changes as $\sim t^{-2/3}$. It changes more slowly than $\sim t^{-1}$, the change in the anisotropy of the model given an isotropic tensor T_α^β.

In all cases, therefore, anisotropy of the tensor T_α^β slows the isotropization of the solution.

We shall now turn to the formulae for the determination of the anisotropy in the RR temperature, anisotropy that is related to the expansion anisotropy. We shall find the red shift for a light ray propagating along the i-axis. Suppose that the scale factor for this axis is $a(t)$. The equation for the photon is then

$$dt = a(t)dx^i. \tag{22.8.6}$$

Let the source and observer be at rest in the reference system, and suppose that they are separated by a coordinate difference Δx^i. Assume that a photon is emitted from the source at t and is received by the observer at t_0. Then we find from equation (22.8.6) that

$$\int_t^{t_0} \frac{dt'}{a(t')} = \Delta x^i. \tag{22.8.7}$$

For a determination of the redshift, it is necessary to find the ratio of the difference in the times of arrival δt_0 of two signals to the difference δt in the times of emission from the source. We find from formula (22.8.7) that

$$\frac{\delta t}{a(t)} = \frac{\delta(t_0)}{a(t_0)}, \quad \frac{\delta t_0}{\delta t} = \frac{a(t_0)}{a(t)}, \tag{22.8.8}$$

since $\Delta x^i \equiv$ const. Hence, the redshift is defined as the ratio of the scale factors for the given direction at the instants of emission and reception of the signals. For the temperature at the instants of emission and reception, we also have

$$\frac{T}{T_0} = \frac{a(t_0)}{a(t)}. \tag{22.8.9}$$

The formulae for the RR temperature anisotropy are obtained in the following way. Suppose that, notwithstanding any deformation anisotropy and before a certain instant of time corresponding to a redshift $z = z_1$, the Universe is transparent to the radiation and the radiation field is isotropic. After the time corresponding to z_1 the radiation propagates freely—without any scattering. Then, for an observer who receives the radiation long afterward, in the directions of the two axes characterized by the scale factors a and b let us say, the difference in the temperature in these two directions is

$$\left.\frac{\Delta T}{T}\right|_{t\to\infty} = \left.\frac{a-b}{a}\right|_{z_1} . \tag{22.8.10}$$

Here $a = b = c$ in the limit $t \to \infty$.

The formulae obtained lead to the following conclusions. If the sharply anisotropic stage ends before the time when $\rho_r = \rho_m$, then, independent of the specific value of this time, the deformation anisotropy will, roughly speaking, be of order unity for the time up to when $\rho_r = \rho_m$. Actually, it follows from formula (22.8.3) that during this stage (when $P = \epsilon/3$), the small deformation anisotropy is "conserved." A more precise estimate of the amplitude of the deformation anisotropy requires an account of logarithmic factors; by doing this one can show that the anisotropy in the deformation rates falls according to a logarithmic law. Subsequently, after the time when $\rho_r \approx \rho_m$ and therefore during the stage when $P = 0$, the anisotropy falls according to the law

$$\frac{a-b}{a} = 6\frac{\rho_{\text{anisot}}}{\rho_m} \sim t^{-2/3} . \tag{22.8.11}$$

Here ρ_{anisot} is the density of the anisotropic neutrino flux. An estimate for $\Delta T/T$ follows from the formulae presented. If one accepts that the intergalactic gas is transparent to the radiation back to the time corresponding to $z_1 = 8$, then one obtains

$$\begin{aligned}\left.\frac{\Delta T}{T}\right|_{t\to\infty} &= \left.\frac{a-b}{a}\right|_{z_1=8} = \left.\frac{6\rho_{\text{anisot}}}{\rho_m}\right|_{z_1=8} \\ &\approx 50\left(\frac{\rho_{\text{anisot}}}{\rho_r}\times\frac{\rho_r}{\rho_m}\right)_{t=t_0} .\end{aligned} \tag{22.8.12}$$

Here t_0 represents the present time, ρ_r is the density of the $T = 3$ K photons, and $\rho_m(t_0) = 10^{-29}$ g cm^{-3} is the baryon density. It follows (see chap. 21) that the density of the anisotropic flux of neutrinos must today constitute a fraction $\rho_{\text{anisot}}/\rho_r \approx 0.1$–$1$ of the background radiation density. For the second ratio, we have $\rho_r/\rho_m \approx 3 \times 10^{-5}$, from which it follows that

$$\frac{\Delta T}{T} \approx 10^{-4} \quad \text{to} \quad 10^{-3} . \tag{22.8.13}$$

Hence, if the expansion anisotropy is large during the stage when the neutrinos cease to interact with the other particles, then, independent of the specific value of the isotropization time of the model, expression (22.8.13) gives the expected value of the RR anisotropy. If the time of isotropization occurs later than when $\rho_m = \rho_r$, one obtains a lower limit on $\Delta T/T$. We emphasize that the presence of ionized intergalactic gas that becomes transparent when $z \approx 8$ is essential in making the estimate (22.8.13).

In HA models with a Euclidean comoving space, according to which formula (22.8.13) was deduced, the radiation anisotropy must be quadrupole in nature. In the direction of the greatest expansion rate, a minimum in the temperature T is observed, while in the direction of the smallest expansion rate (orthogonal to the first direction) a maximum in T is observed. In a measurement from an antenna fixed on the Earth, $\Delta T/T$ must exhibit a 12 hour periodicity.

A review of the measurements of the large-scale RR anisotropy is given by Partridge (1973). The possible upper limit on the anisotropy is $\Delta T/T \lesssim 3 \times 10^{-3}$. This upper limit does not greatly exceed the theoretical estimates given above. Recall once again, however, that these predictions relate only to the model with $\rho_m = \rho_{\text{crit}}$. Concerning models with a curved space, for which $\rho_m \neq \rho_{\text{crit}}$, see the next section.

22.9 The Expected Anisotropy of the Relic Radiation (RR) in Homogeneous, Anisotropic (HA) Models with a Curved Three-Dimensional Space

We shall now turn to the question of the anisotropy of the RR during the early stages of expansion in models with a curved comoving space. Notice first of all the following very important fact. In models with a curved, anisotropic space and during stages close to the Friedmann stage, the deformation anisotropy is "conserved" given the equation of state $P = \epsilon/3$ (Doroshkevich, Lukash and Novikov 1973) in a way similar to that described in the previous section for anisotropic fluxes of neutrinos. According to the conclusions of § 22.7, moreover, after the time of isotropization t_ϕ, the deformation anisotropy decreases only logarithmically during the RD stage. During the entire RD stage up to the moment when $\rho_m = \rho_r$, consequently, the deformation anisotropy is comparable to unity (provided the isotropization does not take place early, whereupon the logarithmic factor would be important). Only after the time when $\rho_m = \rho_r$ does the anisotropy fall according to the $t^{-2/3}$ law. At a certain stage of expansion, the matter becomes transparent to the background radiation, and the photons propagate freely. It is the different expansion rates in the different directions at the moment when the matter becomes transparent that leads to the present-day background anisotropy.

In models with a flat comoving space, the photons moving in the direc-

tions of the principal axes of the rate of deformation tensor always move in the direction of these axes subsequently. As we saw in the preceding section, therefore, the angular distribution of the background intensity over the celestial sphere then corresponds to a triaxial ellipsoidal distribution. The maximum difference in the measured background temperature is thus observed for mutually perpendicular directions on the sky.

It is shown by Novikov (1968) that in models with a curved comoving space the background anisotropy is, generally speaking, not of this nature. For anisotropic models, therefore, there are two most important features of the RR anisotropy. The first is connected with the conservation of the deformation anisotropy during the RD stage. This conservation manifests itself by some kind of anisotropic factor: either by anisotropy in the tensor $T_{\alpha\beta}$ or by anisotropy in the spatial curvature. Moreover, this feature determines the amplitude of $\Delta T/T$. The second feature is related to the motion of rays in the curved, anisotropic, expanding space. It is this feature which gives rise to the distinctive angular distribution on the sky of $\Delta T/T$.

We begin our analysis of the radiation anisotropy with an examination of the background anisotropy for the Bianchi type V model, in which the spatial curvature is isotropic. This analysis will demonstrate the basic characteristics that the curvature of a three-dimensional space introduces into the angular distribution on the sky of $\Delta T/T$. It will also permit one to draw very important conclusions concerning a comparison with observations. For an explanation of the effect of anisotropy in the curvature of the space on the amplitude of $\Delta T/T$, we shall later consider the type VII and IX models.

Consider then the anisotropic type V model with matter at rest relative to the reference system. Remember that the three-dimensional space of this model is a space of constant negative curvature. The curvature of the three-dimensional space does not depend on direction.

The schematic drawing of Figure 59 shows the x-z plane of the three-dimensional space. The solid lines show the paths of photons approaching the observation point O. The dashed lines are in the directions of the z coordinate lines. Finally, the dotted line illustrates the beginning of the propagation of a background photon incident on an observer today; at first such a ray propagates freely, without scattering. It is apparent from the figure that the ray BO (or $B'O$) moves perpendicular to the z-axis at the time when the matter becomes transparent. Inasmuch as the expansion rate differs most strongly from the average in the directions of the x- and y-axes, which are orthogonal to the z-axis, it is clear that the greatest difference in the measured temperatures background is between that in the directions of the rays AO (or in the direction of $A'O$, for which there is an equal background temperature) and BO (or $B'O$). For other

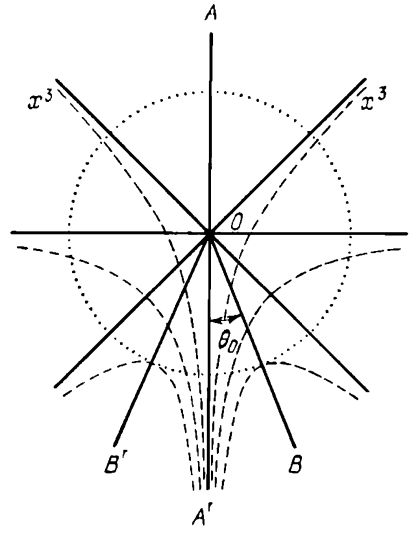


FIG. 59.—The motion of light rays (*solid lines*) in a curved comoving space. The observer is at point 0. The figure depicts the neighborhood of the moment when the Universe becomes transparent to radiation.

planes passing through AOA', the picture is analogous and the temperature anisotropy differs by a factor of $\cos 2\phi$, where ϕ is the angle between the planes. In sum,

$$\frac{\Delta T}{T} = \frac{T - \langle T \rangle}{\langle T \rangle} = Af(\theta) \cos 2\phi \,, \qquad (22.9.1)$$

where A is the constant determining the amplitude of the temperature anisotropy. In order of magnitude, the estimates of A given in previous sections are valid. The function $f(\theta)$ is close to zero everywhere except in a region with a width of order $2\theta_0$, with θ_0 the angle between OA' and OB. The formula for $f(\theta)$ is given below as equation (22.9.3). The angle θ_0 satisfies

$$\sin \theta_0 = \frac{\Omega(1 + z_1)}{2z_1 + \Omega(1 - z_1) - 4z_1(1 - \Omega)[1 + (1 + \Omega z_1)^{1/2}]^{-1}} \,, \qquad (22.9.2)$$

where $\Omega = 8\pi G\rho/3H^2$, H is the average value of the Hubble parameter for $t = t_0$, ρ is the density for $t = t_0$, and z_1 represents the moment when the intergalactic gas becomes transparent.

Thus, the background temperature is almost constant over the whole sky except in a spot whose size is of the order of $4\pi\theta_0^2$ and which encompasses four alternating sectors of excess and deficiency. For the limiting case when the matter becomes transparent at a stage that differs little with respect to deformation from the empty, isotropic Milne model, the function

$f(\theta)$ takes the form (Grishchuk, Doroshkevich, and Novikov 1968)

$$f(\theta) = (1 + \cos\theta)^2 \left\{ \frac{\cos\theta_0}{1 - \cos\theta\cos\theta_0} - \frac{2}{\sin^2\theta}\left[1 + \frac{\cos\theta_0}{1 - \cos\theta_0}(1 - \cos\theta)\right]\right\} . \tag{22.9.3}$$

Consider now the case where the matter in the model moves along the z-axis relative to the system of reference, so that $u_z \neq 0$ (Grishchuk, Doroshkevich, and Novikov 1968). It is easy to see what consequences this leads to. First, the dependence of $\Delta T/T$ on ϕ changes. Instead of equation (22.9.1) for the components of $\Delta T/T$ related to the deformation anisotropy of the model, we obtain

$$\frac{\Delta T}{T} = f(\theta)\,(\sin\alpha\cos^2\phi + \cos\alpha\sin^2\phi) . \tag{22.9.4}$$

Here, $f(\theta)$, is determined by equation (22.9.3), and α depends on the parameters of the model (for more details, see the cited paper). Second, it is necessary to take into account the effect of the velocity of the matter on $\Delta T/T$. From the moment τ'_1, when the matter becomes transparent, to the moment of observation τ'_0, the velocity of the matter relative to the synchronous system sharply decreases. One can therefore neglect the velocity of the observer at the moment τ'_0 by comparison with the velocity of the matter at τ'_1. Thus, in addition to the effect of the background anisotropy considered above, an effect related to the deformation anisotropy and described by equation (22.9.4), there is superposed, in addition, a Doppler effect connected with the motion of the matter along the z-axis. We have assumed that the Doppler effect connected with the motion of the observer is also small compared to the background anisotropy considered above, for which $u_z \equiv 0$. If this is not so, then it is necessary to take into account the motion of the observer too. This leads to an additional dipole effect. For $u_z > 0$, this effect leads to an additional positive $\Delta T_{\mathfrak{D}}$ everywhere inside the spot mentioned above and to a negative $\Delta T_{\mathfrak{D}}$ for the remaining parts of the sky. For $u_z < 0$, the sign of $\Delta T_{\mathfrak{D}}$ inside and outside the spot are reversed. It is clear that for a sufficiently large u_z this effect can become the principal one and can then effectively determine the background anisotropy. For a stage close to the Milne stage, the formula

$$\frac{\Delta T_{\mathfrak{D}}}{T} = \pm (u_z u^z)^{1/2} \frac{\cos\theta - \cos\theta_0}{1 - \cos\theta\cos\theta_0} \tag{22.9.5}$$

describes this effect.

The examples studied above show that the character of the background anisotropy can be complicated. Indeed, the cosmological model examined above predicts a background anisotropy that depends on many parameters. First of all, there are the parameters of the anisotropic model itself. In

addition, three parameters determine the unknown motion of the Sun relative to the radiation field, while three more parameters determine the orientation of the coordinate system with respect to, for example, the galactic coordinate system. One also ought to keep in mind that a quite special model is being considered here; the real number of parameters determining the background anisotropy can therefore be greater still. As an example, one can mention the type V model in which $u_x \neq 0$ and $u_y \neq 0$. In such a model the number of parameters is increased by two. And even more complicated models are also possible. Therefore, a comparison of the predicted background anisotropy with the observations makes for quite a complicated problem.

As a possible method for solving such a problem, one can suggest the following: Several of the observations are first used for a determination of the parameters of a chosen model. This determination is made with the help of a decomposition with respect to a system of orthogonal functions, Legendre polynomials for example. The test of the validity of the chosen model then consists in this case of a comparison of the predicted anisotropy with the remaining observations using parameters already found.

The most important conclusion to be drawn from the considerations of even the quite simple type V model is this. For a test of the possible anisotropy in the expansion of the Universe at an early stage, the isolation of the 12 hour component of the background-radiation anisotropy with respect to a scan of a great circle on the sky is insufficient. (This procedure is sufficient only for the very simple models, the type I models and the simplest type IX models, for example.) Searches for a spot with the predicted $\Delta T/T$ distribution are necessary—notwithstanding the almost complete isotropy of the radiation over the remaining part of the celestial sphere.

We shall now turn to the more general type VII and IX models, which exhibit anisotropy in the three-dimensional spatial curvature (Doroshkevich, Lukash, and Novikov 1974). The curvature anisotropy in these models, as has already been mentioned repeatedly, preserves the expansion anisotropy during the RD stage. It also determines the amplitude of $\Delta T/T$ and plays the same role as do noninteracting particles, which are being disregarded here. Note too that we shall be considering the case in which the matter is at rest relative to the system of reference of the model.

As is shown above in § 22.7, the deformation anisotropy of the model during late stages is practically independent of the initial conditions. Therefore, the amplitude of the RR anisotropy depends only on the time of the beginning of the isotropic stage, t_ϕ, on the moment that the equation of state changes, t_c, and on the instant t_e when the matter becomes transparent. For $\Omega \approx 1$, the amplitude of the RR temperature anisotropy during the anisotropic stage is estimated from those same considerations

that were applied to the type I models in § 22.8. It turns out to be (Doroshkevich, Lukash, and Novikov 1973)

$$\frac{\Delta T}{T} \approx \frac{8}{\ln\left[(t_c/t_\phi)e^8\right]}\left(\frac{t_c}{t_e}\right)^{2/3}. \qquad (22.9.6)$$

A feature of this formula is the extremely weak dependence on the isotropization time t_ϕ. It follows from equation (22.9.6) that if one takes the time of last scattering to coincide with the recombination time and assumes that $\Delta T/T \sim 10^{-3}$, then the instant of isotropization of the model, t_ϕ, is close to the quantum limit of applicability of GTR, $t_{\mathrm{Pl}} \sim 10^{-44}$ s.

The angular distribution on the sky of the RR intensity depends, as is shown above, on the deformation anisotropy along the null geodesics and changes sharply from model to model. In the models considered by us, the simplest structure of the background anisotropy is in the type IX model. In this model, light propagates along the principal directions of the deformation tensor, so that the anisotropy has a quadrupole character. An interesting feature of the type IX model is that the condition $a \approx b \gg c$ holds in the neighborhood of the expansion maximum (see § 22.7). This permits light to successfully circumnavigate the Universe many times in the direction of the axis characterized by the scale factor c during the late stages. The number of circumnavigations is determined by the formula $N \sim (a/c)^{1/2} \sim Q \gg 1$. Up to now it is unclear to what extent astronomical observations limit such a possibility. In type VII models the distribution of the background anisotropy is more complicated. If the observations are carried out for a model whose density is close to the critical density, then the angular dependence of the background anisotropy is close to being quadrupole in character. Subsequently, if $\Omega < 1$, and if the model assumes the Milne behavior, the quadrupole character of the dependence of the background anisotropy is destroyed, and, in analogy to what occurs for the type V model, a spot begins to form.

In distinction to the type V model, the anisotropy in the case under consideration possesses axial symmetry with respect to the axis characterized by the scale factor c (see Fig. 60). During the stage when $t \gg t_e$, the size of the spot is $\langle \Delta\theta \rangle \sim \Omega \ll 1$, while its amplitude is determined by formula (22.9.6). One can find detailed formulae in the paper by Doroshkevich, Lukash, and Novikov (1974).

Hence, what limitations do observations of the degree of RR anisotropy impose upon the parameters of anisotropic cosmological models? Recall that the large-scale RR anisotropy is not greater than 3×10^{-3}. The most important result is that, according to formula (22.9.6), the amplitude of $\Delta T/T$ can be small only in two cases. The first case is where the factor $(t_c/t_e)^{2/3}$ is small. This signifies that the time when the matter becomes transparent is much later than the end of the RD stage: $t_e \gg t_c$. In turn, this is possible only if the amount of ionized intergalactic gas is so signifi-

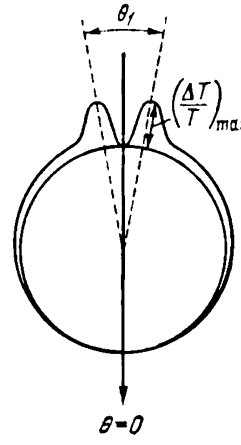


FIG. 60.—Distribution on the sky of the anisotropy of the relic radiation for a model of Bianchi type VII_h.

cant that $\Omega \approx 1$. If the time of transparency t_e corresponds to $z \approx 10$, then $\Delta T/T < 10^{-3}$, and the observational limit to the RR anisotropy agrees with any anisotropic model in which the isotropization occurs not later than the end of the RD stage, i.e., if $t_\phi < t_c$.

If, however, the intergalactic gas density is small and t_e does not differ greatly from t_c, then $\Delta T/T$ can be small only if the denominator is so large that

$$\ln (t_c/t_\phi) \gg 1 .$$

This means that, to obtain an amplitude for $\Delta T/T$ noticeably less than unity, the isotropization must take place extraordinarily early (practically at $t_{Pl} \approx 10^{-44}$ s).

It is finally necessary to emphasize that for $\rho \ll \rho_c$ we cannot reach this conclusion, because $\Delta T/T$ is large only in the small spot and the presence on the sky of a small spot in the $\Delta T/T$ distribution, even a spot with an amplitude near unity, is not at all excluded by present-day observations. To exclude such a possibility, special searches for such a spot on the sky are necessary.

V THE SINGULARITY AND NEW THEORETICAL DEVELOPMENTS

23 THE COSMOLOGICAL SINGULARITY

23.1 Introduction

We now turn to a consideration of an important question in cosmology, the question of the singularity at the beginning of the expansion. To begin, note that a summary of the conclusions reached in the previous parts of the book is that the Universe expands isotropically and homogeneously at least from the time when the $\rho_m \approx \rho_r \approx 10^{-12}$ g cm^{-3}.

What happens earlier? Does the Universe expand in a Friedmannian manner from the singularity, or at least from the Planck time $t_{\rm Pl} \approx 10^{-43}$ s? Or is the earliest epoch substantially non-Friedmannian? Does the matter in the Universe pass through a state of infinitely large density, or at least through a state having the Planck density $\rho_{\rm Pl} \approx 10^{93}$ g cm^{-3}? Or, on the other hand, does the contraction of the Universe from a previous epoch change to expansion for a finite value of the density (see, for example, Alfvén 1971)? It has been suggested there may be an "elementary length" of order $l \approx 10^{-17}$ cm. To this length corresponds a time $\sim 10^{-27}$ s and a density $\rho \approx \hbar/cl^4 \approx 10^{30}$ g cm^3. Perhaps the laws of physics may change during this epoch. We, however, do not believe they do—in contrast, for example, to the ideas of Ginzburg (1975).

According to the Friedmann model, the expansion of the Universe begins from a singularity. Indeed, from the 1930s the question remained whether the existence of a singularity at the start of the expansion is a special property of the Friedmann model and of other sufficiently symmetric models. Does the singularity disappear upon the introduction of small peculiar velocities of the matter or of rotation?

The analogy with the Newtonian problem of the expansion of a sphere reinforces presumptions that it does disappear. If one considers in the Newtonian theory the motion of gravitating particles that simultaneously fly apart radially from one point, then the expansion indeed begins from a singularity.

Given small peculiar velocities in the time-reversed case, however, the points avoid one another near the center, the density is always finite, and a singularity does not develop. Can an analogous situation be possible in the cosmological problem?

Here it is essential to notice a result which was emphasized by Lifshitz and Khalatnikov (1963*a*, *b*). If there is no singularity in the past and if contraction precedes the expansion of the Universe in the past, then the cosmological model describing the passage of matter through the density maximum and describing the subsequent expansion must be stable. In the terminology of Lifshitz and Khalatnikov, the model must apply to the "general solution." In other words, suppose that there is some kind of singularity-free model describing contraction to a finite density and then expansion. Suppose, also, that a small change in the parameters describing the contraction phase of the model leads to the formation of a singularity. It is then evident that this model cannot occur in practice, since random fluctuations leading the solution away from a singularity-free mode always occur. Thus, a solution without a singularity must be nondegenerate and sufficiently general in order to pretend to a description of the real Universe. The conclusions apply to degenerate singular solutions as well. For example, homogeneous isotropic compression is unstable and will be transformed into the stable, general solution.

If, however, the expansion does begin from a singularity, then the requirement on the generality of the solution near the singularity is not obligatory. In this case the initial conditions determining the solution are actually specified by unknown processes that take place at enormous spacetime curvatures, in short, under conditions not describable by present theory. It is possible that processes occurring under these conditions lead to special initial conditions for the expansion of the Universe, to almost complete homogeneity and isotropy, for example (see Peebles 1971*a*). Therefore, if it could be proved that the general solution does not contain a singularity, then this would still not mean, with certainty, that the expansion does not begin from a singularity.

Hence, two different questions have remained: (1) Does there exist a singularity-free, general (in the sense of "stable") cosmological solution? (2) Does a singularity occur in the past, given the conditions that hold in the real Universe?

At the end of the 1960s, a positive answer was given to the second question by Penrose, Hawking, and Geroch. They proved that the expansion of the Universe begins from a singularity if GTR is valid. However, in the approach to the singularity, quantum effects will surely be important and GTR may no longer be valid. They did not, however, establish whether the expansion near the singularity proceeds according to a Friedmann law or to a more complicated law. As a result of their work, the importance of the first question for cosmology diminished sharply. A general analytical solution with a singularity is nevertheless extraordinarily important for the problem of gravitational collapse. Actually, the structure of the solution near the singularity does not necessarily correspond to that of a general

solution, and the problem arises of establishing by some method or other the true character of the beginning of the expansion of the real Universe.

After much effort, Belinsky, Lifshitz, and Khalatnikov (1972) constructed a general (stable) solution with a singularity. In its general properties, moreover, this general solution turned out to be qualitatively the same as the solution near the singularity for the mixmaster model (see § 22.5).

In the subsequent account, we shall discuss the proof that a singularity occurs in the past history of the Universe. We shall also discuss the physical process that may occur near the singularity itself. One can hope that the analysis of these problems and the consequences thereof will, in the future, permit one to establish the true character of the expansion of the Universe during its earliest stages, when the densities are substantially greater than nuclear density.

23.2 The Singularity at the Beginning of the Expansion

Suggestions that deviations from the Friedmann model at the beginning of the expansion do not permit one to avoid the singularity were expressed long ago. They were based on general ideas outlined in the following.

At the present time the radius of the observable part of the Universe is comparable to the gravitational radius. It seemed quite probable, although a strict proof did not exist, that for expansion from inside the gravitational radius, a singularity at the start is inevitable. From this assumption it was suggested that the expansion of the Universe begins from a singularity even if the early stages of expansion are not described by a Friedmann model.

In 1965 Penrose's first theorem concerning gravitational collapse appeared. It was proved in this theorem that a singularity inevitably arises after the contraction of a body to sizes less than the gravitational radius. Following this, there appeared a whole group of theorems by Penrose, Hawking, and Geroch, theorems that concern themselves with the formation of a singularity upon the collapse of an isolated body as well as in the cosmological problem. In 1970, Hawking and Penrose published a generalization of this work. The theorem they formulated makes use of assumptions which are directly verifiable by means of astronomical observations. It is therefore the most suitable upon which to base a discussion of the problem of the singularity in cosmology. One can also find there a detailed bibliography.

All of the studies in the series considered make use of geometrical methods for a proof of the existence of a singularity. Analytic solutions are not constructed near the singularity. They are thus existence theorems; one cannot use them to say anything about the structure of the singularity.

Before passing to an account of the Hawking-Penrose theorem, let us

first discuss the definition of the singularity concept. We are of course interested in true spacetime singularities, i.e., those which exist objectively—independent of the choice of reference system. They are called "true" singularities, in distinction to "coordinate" singularities which are related to an "unfortunate" choice of reference system.

There first of all exist pure coordinate singularities similar to a cross section of the radial lines of a polar coordinate system at the pole. Examples are the cross sections of the geodesic lines forming a synchronous system of reference in a curved spacetime, as has been especially emphasized by Lifshitz and Khalatnikov. On the other hand, it was long ago explained, for example, that the singularity at the gravitational radius of a spherically symmetric Schwarzschild field (see Volume 1) is related simply to the impossibility of continuing a rigid, nondeforming reference system to the interior of the sphere defined by the gravitational radius. If one passes, in this field, to a reference system based on freely falling particles, however, then no singularity at all arises at the gravitational radius.

Even more instructive is the following example (Novikov 1964*a*). Consider in flat spacetime an aggregate of particles at rest at some instant of time. Suppose that the particles begin to move with an acceleration that is inversely proportional to the distance from some point. This acceleration is nongravitational and constant over time for each particle in its proper system of reference. A transition from the spherical coordinates r', θ', and ϕ' of flat space and from the laboratory time t' to the comoving coordinates of the reference system under consideration is specified by the transformations

$$t' = \frac{r}{c}\sinh\left(\frac{gt}{c}\right),$$

$$r' = r\cosh\left(\frac{gt}{c}\right), \qquad (23.2.1)$$

$$\theta' = \theta, \quad \phi' = \phi.$$

Substituting these transformations into the flat-space metric $ds^2 = c^2dt'^2 - dr'^2 - r'^2(d\theta'^2 + sin^2\,\theta' d\phi'^2)$, we obtain an expression for ds^2 in our new system:

$$ds^2 = (gr/c)^2dt^2 - [dr^2 + r^2\cosh^2(gt/c)(d\theta^2 + \sin^2\theta d\phi^2)]. \qquad (23.2.2)$$

The world lines of the particles forming this system are depicted in Figure 61. It follows from equation (23.2.2) that at the center of the system, $r = 0$ (and at this one point only!), there is a singularity; $g_{00} = 0$. The nature of this singularity is moreover apparent from Figure 61. This reference system does not encompass the entire inside part of the light cone whose vertex is at the point 0. Though "only" the one point $r = 0$ is singular in the r, θ, ϕ, t coordinates for all t, this corresponds to the entire interior of the cone $r' - ct' \leq 0$ in the r', θ', ϕ', t' coordinates.

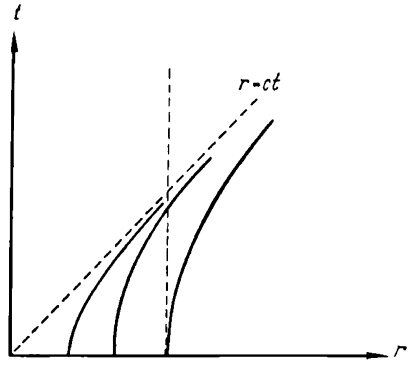


FIG. 61.—World lines of particles at rest in the accelerated coordinate system (r, θ, ϕ const.) of eq. (22.2.1). The dotted vertical line is the world line of a particle at rest in the laboratory coordinate system (r', θ', ϕ' constant).

In the r, θ, ϕ, t system, how can one observe that this system does not encompass the whole of spacetime? One can do this by noticing that a particle that is "falling" freely toward the singular point $r = 0$ reaches this point only after an infinite time as reckoned by an observer at rest in this system. In proper time, however, the particle reaches the singularity after a finite interval of time. We shall now convince ourselves of this. Consider a particle located at $r' = \text{const.}$, $\theta' = \text{const.}$, and $\phi' = \text{const.}$, i.e., at rest in the laboratory system. In the r, θ, ϕ, t system, the equation for its world line is $r = r'/\cosh(gt/c)$, $\theta = \text{const.}$, $\phi = \text{const.}$ The length S of its world line, which defines its proper time t' via $t' = S/c$, is

$$S = \int ds = \int_0^t \frac{(g^2r^2dt^2 - dr^2)^{1/2}}{c^2} = \int_0^t \left[\frac{g^2r^2}{c^2} - (dr/dt)^2\right]^{1/2} dt \qquad (23.2.3)$$

$$= \frac{c^2}{g} \tanh(gt/c) .$$

For $t \to \infty$, the proper time remains finite. The subsequent "life" of the particle after it reaches the point $r = 0$ within a finite proper time t' is not described in the given system. This is so even though the particle world line itself does not encounter any singularities whatsoever here and continues farther (see the dashed line in Fig. 61). Such systems as the t, r, θ, ϕ system are therefore termed "geodesically incomplete," for some geodesics (the ones examined above, for example) are only encompassed partially by these systems. Hence, singularities in geodesically incomplete systems can be removed upon a transition to another system of reference. On the other hand, we are presently interested in true, "invariant," non-removable singularities. At first though, one should mean by a true singularity in spacetime a place where the coordinate-system–independent invariants of the tensor R_{iklm} (if even only one such invariant), the tensor describing the curvature, achieve infinitely large values. At these places the

laws of GTR can be violated. Can one get rid of such a singularity simply by excising a piece of the spacetime around the singularity? Of course, but one ought also to call singular the resulting spacetime, a spacetime with a small hole. Although there are no infinite curvatures in it, the world lines of the particles go right up to the edge of the small hole; thus particles must be "created" or must "disappear" there. One could attempt to exclude the possibility of such small holes by requiring a singly connected spacetime. It is well known, however, that multiply connected spacetimes are possible in which there are no small-hole-type boundaries and which one indeed ought to consider regular from all points of view.

It is most natural to consider a spacetime singularity-free if one can, for all time, continue all of the timelike world lines and null ones as well—into both the past and the future without bound—to an infinite proper length or, for null lines, to an infinite value of an affine parameter. This would mean that the particles move along all paths allowed to them and neither arise nor disappear spontaneously anywhere either in the past or in the future. In such a spacetime there are neither infinite curvatures nor small holes. Such a requirement on the world lines is certainly too strong, however, for it cannot be fulfilled even in flat, Euclidean spacetime.

Thus, referring again to Figure 61, we see that a timelike world line that approaches the direction of a null line (a light trajectory) sufficiently rapidly actually has finite length. A particle moving along such a line, however, experiences an acceleration that tends to infinity. Suppose, for example, that a particle gains velocity according to the law

$$v = c[1 - \cosh^{-4}(gt/c)]^{1/2}, \tag{23.2.4}$$

where g is a constant with the dimensions of acceleration (but is not the acceleration of the particle itself!). Then the proper time t' of the particle that elapses from the start of its motion is

$$t' = \int_0^t [1 - v^2(t)/c^2]^{1/2} dt = \frac{c}{g} \tanh\left(\frac{gt}{c}\right). \tag{23.2.5}$$

For $t \to \infty$, $t' \to c/g < \infty$. Calculations (see Volume 1) then give for the acceleration of a particle in the proper reference system, i.e., the force per unit mass experienced by the particle:

$$a = 2g \frac{\sinh(gt/c) \cosh(gt/c)}{[1 - \cosh^{-4}(gt/c)]^{1/2}} \tag{23.2.6}$$

The acceleration of the particle equals $2^{1/2}g$ for $t = 0$ and tends to infinity for $t \to \infty$. One can now suppose that infinite accelerations are unphysical, that particles cannot move with an acceleration that grows without bound.

Recalling that a timelike geodesic is a world line for a particle moving without acceleration in the proper reference system, we then pass to the following definition: A spacetime is nonsingular if any timelike geodesics,

or null geodesics, can be continued into the past and the future without bound, i.e., to infinite proper length for the timelike geodesics and to an infinite value of an affine parameter for the null geodesics. Such a spacetime is termed "causally, geodesically complete." The requirements on the completeness are the minimum necessary so that the spacetime does not contain a singularity. It is necessary to point out that a spacetime not satisfying these requirements, however, one with a singularity, does not necessarily contain points with infinite curvature or with a small hole. Examples have been constructed in which a timelike geodesic rapidly approaches a null geodesic yet has a finite proper length.

From the physical point of view, of course, one ought to take as singular any spacetime in which the geodesic world line of a particle cannot be continued without bound with respect to the proper time of this particle, for such a singular spacetime would lead to a violation of conservation laws. We shall confine ourselves to this definition.

As applied to the cosmological problem, the Hawking-Penrose (Hawking and Penrose 1970) theorem reads as follows:

THEOREM. *A spacetime M cannot be causally, geodesically complete if the GTR equations hold and if the following conditions are fulfilled:*

1. The spacetime M does not contain closed timelike lines.
2. The conditions

$$\epsilon + \sum_{\alpha} p_{\alpha} \geq 0 ; \qquad \epsilon + p_{\alpha} \geq 0 , \quad \alpha = 1, 2, 3 , \tag{23.2.7}$$

on the equation of state are fulfilled, where ϵ is the energy density and p_1, p_2, and p_3 are the three principal values of the pressure tensor.

3. On each timelike or null geodesic, there is at least one point for which

$$K_{[a}R_{b]cd[e}K_{f]}K^{c}K^{d} \neq 0 , \tag{23.2.8}$$

where K_a is the tangent to the curve at the given point and where the brackets on the subscripts imply antisymmetrization.

4. The spacetime M contains either (*a*) a point P such that all diverging rays from this point begin to converge if one traces them back into the past, or (*b*) a compact spacelike hypersurface.

Let us now discuss the conditions of the theorem. Requirement (1) is natural. The presence of closed timelike lines would violate causality; indeed, the existence of such lines would seem by itself to signify a certain kind of "singularity." Hawking and Penrose (1970), however, emphasize that Hawking (1966*b*) has proved a theorem concerning the existence of a singularity in an expanding closed Universe without conditions relating to closed timelike lines. But it was required there that all of the matter be expanding on some compact spacelike surface. Such a theorem is therefore not formally applicable to the Universe, in which there are collapsing

bodies even if one takes the Universe to be closed. This subtle requirement is essential for the mathematical proof; to an astrophysicist, though, the theorem is a very strong argument that the formation of closed timelike lines will not eliminate the inevitability of singularity.

Requirement (2) is always fulfilled for all known forms of matter and fields under conditions when the spacetime curvature is far from the Planck value of $(10^{-33}\ \text{cm})^{-2}$ and when quantum effects are not important.

Requirement (3) is purely mathematical. By all accounts, it is certainly valid in any nondegenerate solution. Nevertheless, the occurrence of a geodesic with such degenerate properties that equation (23.2.8) fails all along it can still hardly eliminate the inevitability of a singularity.

Finally, requirement (4) in essence represents the global physical conditions that lead to the appearance of a singularity. These conditions suggest that the gravity of all the matter is so strong that a singularity is inevitable. Requirement (4*a*) is in principle subject to direct verification by observations. In reality, however, this requirement is tested through a combination of observational data and very simple calculations. Suppose, for example, one proves that this condition is fulfilled in a Friedmann model and then shows, in agreement with observations, that the Friedmann model applies to the past to the time when the rays again begin to converge. Then one has proved that requirement (4*a*) is fulfilled.

We have already considered the problem of the convergence of rays in the past (§ 3.3). Given a homogeneous distribution of matter and the condition $\Omega_0 = 1$, the rays begin to converge at a redshift of the order of $z \approx 1$, as the formulae of § 3.3 show. But even if we assume that the average density is $\Omega_0 = 0.03$, that all the matter is in galaxies, and that light rays that do not encounter a single galaxy do not converge at a value of $z \approx 1$, then, on passing to still larger values, $z > 10$–20, we still reach conditions when any light rays must converge; at this time, moreover, the Universe is expanding according to a Friedmann law. Thus, requirement (4*a*) is fulfilled in the real Universe.

Consequently, the expansion of the Universe begins from a singularity—in any case, this is so if GTR is valid. Unfortunately, the Hawking-Penrose theorem tells us practically nothing about the structure of the singularity. Strictly speaking, it does not even say whether the singularity is in the past or in the future! (For details, see the original paper by Hawking and Penrose 1970.) It simply maintains that at least one geodesic world line of a particle or a photon cannot be continued without bound into the past or the future. Notwithstanding possible exotic examples wherein the spacetime is geodesically incomplete in just this way but nevertheless lacks infinite curvature, almost all specialists assume that in the general, nondegenerate case, a singularity signifies an infinite spacetime curvature.

Two questions remain unanswered by the theorem. First, does all the material go through a singular state in the past, or only a part of it? Second,

how does the expansion proceed near the singularity itself? The second of these questions is analyzed in the next section (see also Trautman 1973).

23.3 The General Cosmological Solution with a Singularity

We have already emphasized in § 23.1 that an examination of the cosmological singularity in the past, at the start of the expansion, produces no special basis upon which to assume that the character of the solution is described by the most general solution rather than by a certain special, degenerate solution. The character of the expansion in this case is determined by the initial conditions at the singularity, conditions that we do not know and conditions that can determine this character according to some special solution that is not the most general. We shall consider one such possibility later, that of the intense creation of particle-antiparticle pairs near the singularity. This leads to an isotropic expansion. But of course the solution describing the most general expansion near the singularity is of enormous interest as an aid to understanding what can occur and, in the final analysis, as an explanation of what in reality does occur.

Besides this we ought to emphasize that it is indeed the general solution that describes collapse, the contraction to a singularity, both for cosmological models (if the expansion changes to contraction, i.e., if $\rho > \rho_{\text{crit}}$) as well as for models of the collapse of an isolated body that contracts within its gravitational radius (see Volume 1).

The general solution near a singularity has been constructed by Belinsky, Lifshitz, and Khalatnikov (1972). In this section the results of that study are presented. It turns out in the general case that, upon the approach to the singularity, the solution describing the deformation has locally, i.e., in the neighborhood of every point, the same type of character that occurs in the Bianchi type VIII or IX models. The solution consists of alternating Kasner epochs and describes an oscillatory regime of approach to the singularity.

For convenience in this outline, we shall in this section be describing, as we did in § 21.5, contraction or collapse. For a description of expansion, it is merely necessary to change the direction of the flow of time. We now proceed with the construction of the general solution.

The so-called generalized Kasner solution appeared early in Lifshitz and Khalatnikov's (1963*a*, *b*) work. This solution describes the deformation locally in the same way as does the Kasner solution (see equations [19.3.6] and [19.3.7]). In the generalized case, however, the direction of the principal axes of the deformation and the size of the Kasner exponents p_1, p_2, and p_3 change from point to point in the three-dimensional space. The metric of the generalized Kasner solution is therefore written in a synchronous reference system in the form

$$g_{\alpha\beta} = a^2 l_\alpha l_\beta + b^2 m_\alpha m_\beta + c^2 n_\alpha n_\beta ; \qquad (23.3.1)$$

$$a \sim t^{p_l}, \quad b \sim t^{p_m}, \quad c \sim t^{p_n}; \tag{23.3.2}$$

$$p_l + p_m + p_n = p_l^2 + p_m^2 + p_n^2 = 1. \tag{23.3.3}$$

The Greek indices run over the values 1, 2, and 3. The exponents p_l, p_m, and p_n are functions of the spatial coordinates, as are the spatial unit vectors l, m, and n. We now write the Einstein equation with respect to the synchronous reference system and with the speed of light set equal to unity:

$$-\frac{\partial}{\partial t}\mathfrak{D}_\alpha{}^\alpha - \mathfrak{D}_\alpha{}^\beta\mathfrak{D}_\beta{}^\alpha = 8\pi G(T_0{}^0 - \tfrac{1}{2}T); \tag{23.3.4}$$

$$\mathfrak{D}_\alpha{}^\beta{}_{;\beta} - \mathfrak{D}_\beta{}^\beta{}_{;\alpha} = 8\pi G T_\alpha{}^0: \tag{23.3.5}$$

$$-\frac{1}{\sqrt{g}}\frac{\partial}{\partial t}[(\sqrt{g})\mathfrak{D}_\alpha{}^\beta] - R_\alpha{}^\beta = 8\pi G(T_\alpha{}^\beta - \tfrac{1}{2}\delta_\alpha{}^\beta T). \tag{23.3.6}$$

Here $\mathfrak{D}_{\alpha\beta} = \frac{1}{2}\partial g_{\alpha\beta}/\partial t$, while the semicolon denotes covariant differentiation in the three-dimensional space with the metric $g_{\alpha\beta}$. Finally, $R_\alpha{}^\beta$ is the three-dimensional Ricci curvature tensor constructed from $g_{\alpha\beta}$.

We shall be studying the solution of the Einstein equations near the singularity. Estimates show that in this case one can neglect the components of the stress-energy tensor in equations (23.3.4) and (23.3.6) by comparison with the terms on the left-hand side of these equations (the vacuum stage). We shall here not be considering equation (23.3.5). Note though that it allows one to calculate the distribution and motion of the matter once the solution for the $g_{\alpha\beta}$ has been found. Hence, let us study equations (23.3.4) and (23.3.6) with their right-hand sides equal to zero.

If in equation (23.3.6) one neglects $R_\alpha{}^\beta$, then the solution of equations (23.3.4) and (23.3.6) will be the generalized Kasner solution (23.3.1–3). We shall assume that during some interval when t changes, the solution (23.3.1–3) is indeed valid, so that one can neglect the $R_\alpha{}^\beta$. We shall then continue the solution to the singularity ($t \to 0$), and shall determine the boundary of the region of applicability of equations (23.3.1–3). To do this, we shall for each point of the three-dimensional space project all of the tensors in the directions of the vectors l, m, and n. In order that solution (23.3.1–3) satisfy equations (23.3.4) and (23.3.6), it is necessary that the diagonal projections of the Ricci tensor be less than the remaining terms in equation (23.3.6); these terms are of order t^{-2}. Therefore, the conditions

$$R_l{}^l < t^{-2}, \quad R_m{}^m < t^{-2}, \quad R_n{}^n < t^{-2} \tag{23.3.7}$$

must be fulfilled. Concerning the nondiagonal components, the original Belinsky, Lifshitz, and Khalatnikov (1972) analysis shows that one can always neglect them to first order upon a continuation of the solution to the singularity if one can neglect them at first. It is necessary to keep in mind always that we are studying the general solution. Degenerate cases are, of course, always possible, but they require a special choice of the initial conditions. We shall not discuss them now.

To first order, therefore, we must examine four equations: equation (23.3.4) for the 0–0 components and equation (23.3.6) for the three diagonal projections l–l, m–m, and n–n. The diagonal projections of the Ricci tensor, $R_l{}^l$, $R_m{}^m$, and $R_n{}^n$, yield terms of the form

$$A = \left[\frac{l\cdot\nabla \times (al)}{bcl\cdot m \times n}\right]^2 , \tag{23.3.8}$$

and other such terms with a cyclic permutation of the vector l with m and n and with corresponding changes in the scalar functions a, b, and c.

In the generalized Kasner solution, when $t \to 0$, one of the functions a, b, and c grows, while the other two fall. Suppose that a is the growing function. The substitution of equations (23.3.2) and (23.3.3) into expression (23.3.8) shows that this expression grows as

$$A \sim t^{-2(|p_l|+1)} , \tag{23.3.9}$$

where t decreases. The quantity in the parentheses in the exponent of formula (23.3.9) is greater than unity. It grows more rapidly than does t^{-2} when $t \to 0$. At some instant of time, therefore, the term A becomes equal to the remaining terms in equation (23.3.6), terms which grow as t^{-2}, and the Kasner solution breaks down.

The four equations—the 0–0, l–l, m–m, and n–n projections mentioned above—which are the only important equations to first order, take on the following form in this epoch of Kasner breakdown:

$$\begin{aligned}
&a^{-1}\frac{d^2a}{dt^2} + b^{-1}\frac{d^2b}{dt^2} + c^{-1}\frac{d^2c}{dt^2} = 0 ,\\
&(abc)^{-1}\frac{d}{dt}\left(bc\frac{da}{dt}\right) + \frac{\lambda^2a^2}{2b^2c^2} = 0 ,\\
&(abc)^{-1}\frac{d}{dt}\left(ac\frac{db}{dt}\right) - \frac{\lambda^2a^2}{2b^2c^2} = 0 ,\\
&(abc)^{-1}\frac{d}{dt}\left(ab\frac{dc}{dt}\right) - \frac{\lambda^2a^2}{2b^2c^2} = 0 .
\end{aligned} \tag{23.3.10}$$

Here $\lambda = (l\cdot\nabla \times l)/(l\cdot m \times n)$. These equations coincide with the corresponding equations for the Bianchi type IX model (see § 22.4) for the period when one Kasner epoch changes into another. The only difference is that in equations (23.3.10), λ is a function of the spatial coordinates. For the local properties of the solution, however, this is of no significance. Therefore, the exchange of Kasner epochs takes place in exactly the same way as it does for the mixmaster model. In sum, the evolution as t approaches zero takes place locally, at each point, in qualitatively the same way as it does in the mixmaster model described in § 22.2.

One ought to add only that when account is taken of the nondiagonal projections of equations (23.3.6), the following conclusion is reached.

Upon an exchange of Kasner epochs, the vectors l, m, and n rotate at each point as well. The same type of rotation also takes place in the case of the homogeneous type IX model when matter moves relative to the reference system of the model.

Locally, therefore, the solution describing the metric in the general case has an oscillatory character at each point of the space as the singularity is approached. Such behavior is depicted in Figure 57. To describe expansion from the singularity, it is merely necessary to change the direction of the time flow.

In concluding this section, we emphasize once again the point that the beginning of the expansion of the Universe is not necessarily described by the general solution. This point is discussed in detail at the beginning of the section.

24 PHYSICAL PROCESSES NEAR A SINGULARITY AND DEVELOPMENTS IN THE THEORY OF GRAVITATION

24.1 Introduction

GTR leads to the conclusion that in the past history of the Universe there was a singular state. This conclusion is based on the classical GTR equations and on definite ideas concerning the nature of the terms on the right side of these equations (i.e., the terms that describe the behavior of the matter that fills space).

The existence of a singular state apparently means that infinite values of the spacetime curvature and an infinite matter density are achieved, at least formally. Given such a state, it becomes impossible to continue a solution any further into the past.

In the previous chapter, we emphasized that a more fashionable and more precise definition of a singularity consists of this statement: Given a singularity, the world lines of particles cannot be continued beyond a certain point; moreover, the world lines so restricted include null world lines, the trajectories of particles with zero rest mass.

But there exist in physics well-known charge conservation laws, which are formulated in the language of particles to express both the continuity of particle world lines and the impossibility that such lines can begin or end somewhere. Physics therefore requires that singularities be avoided or that our conception of them be changed. On the other hand, when the curvature and the density approach infinity, it is wholly probable and possible that the original propositions of the analysis (GTR and the properties of matter) are changed. Thus, when considering behavior near a singularity, one can expect to go beyond the bound of those ideas which were used earlier.

The distinction between the processes considered earlier and those to be discussed here is based on the degree of reliability. Our discussion of electron-positron annihilation and the readjustment of temperature, of the weak interactions which maintain the neutron-proton ratio in equilibrium, of nuclear reactions—all these were based on well known physical laws verified in the laboratory. Now, in contrast, we collect together some speculative and highly uncertain problems. We will first of all consider the statistical properties of a plasma containing many nucleons and antinucleons under various assumptions relating to the strong interactions and the excited

states of the nucleon (Hagedorn's theory and Omnès's theory). We shall then look at specific quantum phenomena which occur in a gravitational field. The necessary consequences of the experimentally established fact that various fields (the electromagnetic field, the neutrino field, and so forth) are quantized shall be considered first. The metric and the geometry of spacetime will be considered classical and predetermined. More precisely, the small-scale, fluctuating part of the metric, the part that constitutes gravitational waves, is examined in a quantum form, together with the electromagnetic and other fields, on a background determined by an average metric.

We shall next study the back reaction of the quantum fields on the metric. Under definite conditions, one can regard this influence as corrections to the classical GTR equations.

In § 24.6, we shall briefly examine the quantization of the Universe as a whole. To do this, we consider the possible evolutionary paths of the Universe as trajectories having various probabilities or, more precisely, having various complex probability amplitudes. Now Wheeler has introduced the concept of a superspace of trajectories, while Misner has examined mini-superspace, which confines itself to trajectories of homogeneous models of the Universe. Since, however, the Universe is so vast—even if it is closed, it contains on the order of 10^{80} baryons and 10^{90} photons—such a global approach is unphysical if local effects are ignored.

It is natural to pose the question of whether, in addition, physical laws can change near a singularity. The problem of the nature of the state of the Universe near a singularity is still very far from final resolution. Diametrically opposite hypotheses are therefore possible and do indeed exist. Along with the charge-symmetric hypothesis described above (and the charge-symmetric hypothesis of Alfvén and Klein and of Omnès), a 100% charge-asymmetric cosmological hypothesis is also possible. In this asymmetric case, what is considered is a cold initial state for the Universe and a unique spectrum of perturbations of the geometry and of the density. The long-wavelength part of the spectrum is then responsible for galaxy formation, while the short-wavelength part of the spectrum gives acoustic waves. Upon decaying, these waves create entropy.

In this chapter the examination of phenomena near the singularity serves as a basis for a modification of classical theories. In furtherance of this philosophy, we also examine in this chapter (§§ 24.9, 24.10) several other assumptions that go beyond the bounds of standard theoretical physics, even if they are not directly connected with the existence of a singularity. The steady-state and C-field theories developed by Hoyle and his coworkers relate foremost of all to these assumptions (§ 24.9). The so-called Mach principle—the hypothesis concerning a direct effect of the entire Universe on every local phenomenon—is also examined. Dirac perceives

concrete manifestations of this connection through numerical coincidences between local physical constants and quantities characteristic of the Universe as a whole.

Several generalizations of GTR are studied in § 24.11, as is the possible cosmological significance of these generalizations. Exotic models of the Universe with strange topology are considered in § 24.11. It is curious both that there is a plethora of such models and that it is difficult to make a choice between the standard and exotic models based on observational criteria. In § 24.12 we then remark upon the local topological structure of spacetime.

Taken together, the contents of this chapter correspond to Il'f and Petrov's[1] formulation: The contents constitute "everything that is required for the future."

For further references to the problems to be discussed here, see Wheeler (1968), Ginzburg (1971), Weinberg (1972), Sciama (1971), Kirzhnits and Linde (1974), Zel'dovich, Kobzarev, and Okun' (1974).

24.2 Cosmological Consequences of Hagedorn's Theory

Hagedorn's theory (see § 6.4) is characterized by these distinctive features: (1) a maximum temperature $kT_{\max} \approx m_\pi c^2$, so that it is of the order of $T_{\max} \approx (1\text{-}2) \times 10^{12}$ K; and (2) a "soft" equation of state, $p \ll \epsilon$, for temperatures approaching the maximum. For temperatures lower than $T_{\max}$, however—e.g., for $T < 5 \times 10^{11}$–10^{12} K—the thermodynamic properties of this theory do not differ from those of the ordinary theory, which is based upon the existence of a finite and small number of kinds of particles.

The first feature is important with regard to the question of the number of quarks or partons remaining after expansion of the Universe causes cooling. If one takes the quark mass to be m_q, then the fraction of the free quarks in equilibrium at $T_{\max}$ with respect to the total number of all particles is of the order of $\exp[-m_q c^2/(kT_{\max})] \sim \exp(-m_q/m_\pi)$. For a sufficiently large quark mass, consequently, one satisfying $m_q > 50m_\pi$ or $m_q > 7m_p$ (m_q is the quark or parton mass, m_π is the pion mass, and m_p is the proton mass), the initial equilibrium number of quarks is less than that predicted for the period after the freeze-out by ordinary particle theory. Frautschi, Steigman, and Bahcall (1972) have especially stressed this point.

If accelerator or cosmic-ray experiments prove that stable, heavy, free quarks exist, then Hagedorn's theory would appear to be the anchor saving contemporary cosmology. As was already remarked, however, stable, heavy

1. Il'f (1897–1937) and Petrov (1903–1942) were famous Russian humorists. They are the authors of the comic novels, *Twelve Chairs* and *Golden Calf* about life in the USSR before World War II and, also, of a travel book, *One Store America* (1938) which is full of good humor and warmth.

quarks have not been observed. But the absence of cold quarks in such a situation is not a confirmation of Hagedorn's theory. Consequently, the aforementioned prediction of Hagedorn's theory remains in the air. Notice also that Hagedorn's theory agrees poorly from a logical viewpoint with the quark hypothesis. For if baryons consist of quarks, then one cannot regard the resonances, excited stages of the baryons, as independent, noninteracting particles.

Turn now to the second feature. A change in the equation of state near the singularity would somewhat change the expansion law, and thus the dependence of the density and the radius on time. Given the fixed, present-day values of the density and the Hubble parameter, however, all of the changes are strictly confined to the period when $T \sim T_{\max}$, so that there are no changes whatsoever during the stage of nucleosynthesis and recombination.

Can Hagedorn's theory change the picture of the development of perturbations? In the thermodynamic theory of equilibrium perturbations, the pressure and the elasticity ($\partial p/\partial \rho = b^2$, where b is the sound velocity) are important. In particular, the pressure gradient $\partial p/\partial x = (\partial p/\partial \rho)(\partial \rho/\partial x) = b^2 \partial \rho/\partial x$ prevents the emergence of density fluctuations. For a specified energy of each acoustic mode in thermodynamic equilibrium equal to kT, the corresponding deviation of the density from the average value is inversely proportional to the sound speed. Consequently, thermodynamic fluctuations in Hagedorn's theory are substantially greater (in the limit $\rho \to \infty$, by an infinite factor) than those in the theory of noninteracting particles.

In cosmology, however, the conditions necessary for obtaining thermodynamic equilibrium among fluctuations during the early stage do not exist.

A more natural formulation of the question consists of a specification of the initial perturbation of the metric. Recall that according to Lifshitz's solution, a metric perturbation remains constant near the singularity, going neither to infinity nor to zero. Moreover, the period during which this behavior occurs extends until such time as $bt < \lambda$ and $\delta\rho/\rho < 1$. Since $b < c$, a metric perturbation is in any case constant as long as there is no exchange of information between neighboring regions (those of minimum and maximum density, for example). This result was previously formulated in the following way: during this period, separate parts of a perturbed Friedmann solution develop independently, with the metric preserved. The extreme estimate $ct < \lambda$, with the velocity of light replacing the velocity of sound, is valid for any equation of state.

Let us now find the characteristic wavelength at the instant when the Hagedorn equation of state turns into the ordinary equation of state $P = \epsilon/3$. At this time, $T \approx 10^{12}$ K, $t \approx 10^{-4}$ s, $n_B \approx 3 \times 10^{28}\ \mathrm{cm}^{-3}$, $N = n(ct)^3 \approx 10^{48}$, and $M = 10^{-8} M_\odot$. The mass of a perturbed region is in-

significant! But for any larger mass, the metric perturbations survive without changing until the period when all of the resonances annihilate and decay, i.e., until the effects characteristic of Hagedorn's theory disappear. Now questions concerning galaxy formation and fluctuations of the relic temperature are just those questions for which the perturbation theory is connected with observations. For such phenomena, then, one can be certain that Hagedorn's theory does not change the conclusions drawn earlier. The softness ($P \ll \epsilon$) of Hagedorn's equation of state is important for problems of the formation of primordial black holes and also for their accretion of surrounding material. These topics are discussed later.

24.3 Quantum Phenomena That Take Place near Singular States of a Metric and in Strong Gravitational Fields

It has been remarked above that under extreme conditions it is necessary to simultaneously take account of GTR and of quantum effects. We shall find that the account of quantum effects will introduce important theoretical changes in the deductions from classical GTR.

Can one expect important effects under all conditions? To answer this, we note that GTR does not bring new physical quantities into the theory. The cosmological constant Λ, if it differs from zero, is important only over very large scales, for which there are no quantum effects. We shall moreover make use of the famous constant h introduced by Max Planck into the theory of radiation in 1899 [though it is now customary to use the quantity $\hbar = h/(2\pi) = 1.05 \times 10^{-27}$ g cm^2 s^{-1}]. He understood clearly the significance of the quantization idea as a concept applying to all of physics, indeed to all of natural science. Considering c, G, and $\hbar$ as three equally fundamental quantities, Planck showed that quantities of any dimensionality can be expressed in terms of them alone. In particular, one can express in terms of c, G, and $\hbar$ a unit of length l_{Pl}, a unit of time t_{Pl}, a unit of mass m_{Pl}, and a unit of density ρ_{Pl} as follows:

$$l_{\mathrm{Pl}} = (G\hbar/c^3)^{1/2} \approx 1.6 \times 10^{-33} \text{ cm} ;$$

$$t_{\mathrm{Pl}} = l_{\mathrm{Pl}}/c = (G\hbar/c^5)^{1/2} \approx 5.3 \times 10^{-44} \text{ s} ;$$

$$m_{\mathrm{Pl}} = (\hbar c/G)^{1/2} \approx 2.2 \times 10^{-5} \text{ g} ;$$

$$\rho_{\mathrm{Pl}} = m_{\mathrm{Pl}}/l_{\mathrm{Pl}}{}^3 = c^5/(G^2\hbar) \approx 5 \times 10^{93} \text{ g cm}^{-3} .$$

It is useful to notice the similarity between the Coulomb law $V = e^2/r$ and the Newtonian gravitational law $V = Gm^2/r$. Since e^2 and Gm^2 have the same dimensionality, $Gm^2/(\hbar c)$ is evidently a dimensionless quantity similar to the fine-structure constant $e^2/(\hbar c) = \alpha = 1/137$. For the electron and the proton, for example, we have $Gm_e{}^2/(\hbar c) \approx 2 \times 10^{-45}$ and $Gm_p{}^2/(\hbar c) \approx 6 \times 10^{-39}$. Note that the condition $Gm^2/(\hbar c) = 1$ gives the characteristic mass m_{Pl} presented above. The length l_{Pl} is the "Compton wave-

length" for the mass m_{Pl}; namely, $l_{Pl} = \hbar/(m_{Pl}c)$. Finally, note that another convention is used in elementary-particle theory. There we take $\hbar = c = 1$. In such a system of units, the length and time have the same dimensionality, inverse to the dimensionality of mass: $[l] = [t] = [m^{-1}]$. The product Gm^2 is dimensionless, so the dimensions of G are $[m^{-2}] = [l^2]$; corresponding to G is an "area" or "cross section," roughly equal to 2.5×10^{-66} cm^2.

These quantities characterize the conditions under which quantum gravitational effects play the principal role. As an example, the conditions require that the spacetime curvature $R_{\beta\delta}{}^{\alpha\gamma}$ be of the order of $\sim l_{Pl}{}^{-2} \sim 10^{66}$ cm^{-2}. Such a curvature can arise in a vacuum but is certainly not a "necessary" attribute of a vacuum. On the other hand, if a matter density achieves a value of the order of ρ_{Pl}, then the corresponding curvature that follows from the GTR equations is of the order of $l_{Pl}{}^{-2}$; in this sense such a curvature is "necessary."

As simple as it may be to determine the conditions under which quantum phenomena are important, it is nevertheless quite difficult to calculate just what takes place. (See DeWitt and Wheeler 1968; Ginzburg, Kirzhnits, and Lyubushin 1971.) Indeed, it becomes difficult even to formulate the problem since ordinary physics, quantum physics included, is considered in the realm of a specified spacetime manifold. In quantum physics, classical trajectories and fields are replaced by the concept of wave functions, with the help of which one can express the predicted probabilities for experimental results. However, the coordinates and time are still considered as ordinary, determinate quantities (C-numbers). The curvature of spacetime does not change any relevant theoretical aspects if this curvature is less than $l_{Pl}{}^{-2}$. In a quantum-gravitational regime, however, space and time themselves might well acquire probabilistic, indeterminate properties.

In cosmology one method for dealing with this problem consists of posing questions (and calculating quantities) relating to that period when the Universe is no longer in the singular state—when both the curvature and the matter density are nowhere huge. Such an approach would be similar to that of the S-matrix theory. As is well known, Heisenberg suggested that one consider only the states before and after elementary-particle collisions and refrain from giving a detailed description of the collision act itself. The value of such an approach is that it allows one to prove that, in principle, a result does indeed exist given a definite initial situation. This is, however, insufficient if one wishes to obtain the specific properties of that result. Many important and specific results have been obtained in S-matrix theory, but even in the realm of elementary-particle phenomena the theory is insufficient to answer all questions one can pose. Characteristically most recent progress is connected with the Lagrangian approach, not to the S-matrix one!

A quantum gravitational theory is especially necessary in cosmology inasmuch as it is certain that the Universe (one can apparently strengthen

this to the whole Universe and all of the matter in the Universe) passes through a state, an analysis of which requires the use of this theory. As we saw above, moreover, such an examination is necessary because of the large variety of classical (nonquantum) cosmological solutions. Perhaps a quantum gravitational theory of the singular state will indicate how to choose from this large number of possibilities.

A complete quantum-gravitational, cosmological theory does not at present exist. There are only particular results, some of which are set forth below. However, even given such an uncertain form for the theory, one can discern indications that perhaps an anisotropic singular metric will turn out to be forbidden and that only the quasi-isotropic solution will be allowed (see Zel'dovich 1970*c*; Lukash and Starobinsky 1974). An approach to explaining the entropy of the Universe is also contemplated. Ideas on the formation of primeval black holes near the singularity and their subsequent quantum evaporation, with effective baryon nonconservation, are now very fashionable speculation. In short, the problem of a quantum-gravitational theory is doubtless of enormous significance for theoretical cosmology and, indirectly, through a chain of inferences, for observational cosmology too.

As we proceed with the discussion of this theory, we shall set forth hypotheses and questions underlying the research along with the firmly established facts. So let us not waver; we devote the following sections to an analysis of quantum-gravitational theory.

Quantum electrodynamics serves as an analog for such a theory. Using it, physicists have managed to obtain remarkable agreement with experiment for certain specific effects predicted by the theory. We refer to the Lamb shift of the levels of the hydrogen atom and the anomalous magnetic moment of the electron. These advances were achieved by means of a systematic application of the quantum theory.

Indeed, this required the introduction of such new concepts as mass renormalization, charge renormalization, and vacuum polarization. However, there has been no need to introduce an elementary length or to repudiate the general principles of quantum mechanics. To be sure, quantum electrodynamics is an inspiring example for a future quantum gravitational theory.

In a series of studies the logical scheme of such a theory is analyzed and the quantum gravitational corrections to quantities observable in laboratory experiments are computed. The first step, taken in the 1930s, was the quantization of the linear theory of gravitational waves. In this case gravitational waves were considered as small perturbations of a flat-space geometry or as an external (not a geometrical) tensor field inserted into flat space. From today's point of view, the results are straightforward. The quanta of this field are called gravitons; the graviton energy is $\hbar\omega$, they are bosons with spin 2 and with zero rest mass. In subsequent work, the non-

linearity of the original classical theory (GTR) was found to be important. Indeed, gravitons themselves possess energy and momentum—although their rest mass equals zero—and are, consequently, gravitational field sources themselves. Feynman (1963) began the work leading to a consistent account of this fact while Fadeev and Popov (1967) and De Witt (1967*a*, *b*) more recently shed new light on this issue. Now specifically quantum-gravitational effects are small in laboratory physics and even in astronomy if one excludes the singularity. Thus, Feynman's activity and that of several other authors was inspired more by esthetic goals, a feeling, moreover, that Feynman did not hide.

In cosmology the situation is quite different. For $t \approx t_{\mathrm{Pl}}$, quantum-gravitational effects are of order unity, so there is much interest in obtaining even a rough idea of their character. As will be shown below, the most important effect is probably the creation of particles or of pairs of particles in strong gravitational fields.

The effect of a gravitational field on the motion of particles and on the propagation of fields can be completely described by the specification of the spacetime metric. In particular, the constant G does not enter the equations governing particle motion and wave propagation in a specified spacetime. The quantity G appears only in the converse problem of the effect of the matter on the metric.

One can obtain the most general ideas that apply to the process of creation by beginning with an examination of a classical (nonquantum) linear wave. In flat spacetime, a wave propagates in such a way that its energy and frequency are separately conserved. Given a curved and nonstationary metric, this is not so, but the important limiting case of geometrical optics exists if the wavelength and the period of the wave are small by comparison respectively with the size of a region over which deviations of the metric from Euclidean geometry are noticeable and with the time characterizing a noticeable change in the metric. Now geometrical optics encompasses two concepts. The first is the concept of a ray, which is to a wave packet what the concept of trajectory is to a particle. The second is the concept of adiabatic invariance, which relates to the amplitude and intensity of a wave field. According to this concept, the energy of a wave field changes in proportion to its frequency. Consequently, the ratio of the energy to the frequency is an adiabatic invariant, remaining constant within the realm of geometrical optics. However, this ratio is just proportional to the number of field quanta: $E = n\hbar\omega$, $E/\omega = n\hbar$. Therefore classical geometrical optics encompasses the notion of conservation of the number of quanta even though no quantum effects at all are considered in this theory. This aspect of the connection between the classical and quantum theory is considered in Paradoksov's (1966) article, "How Quantum Mechanics Helps One to Understand Classical Mechanics." See Zel'dovich (1965*a*) for the general expression giving the number of photons in an arbitrary electromagnetic field.

But, upon a rapid change in the metric, the adiabatic invariance is destroyed. The number of quanta therefore changes; quanta are created or destroyed. It is important that the change in the number of quanta occurs in the absence of sources of the fields such as moving charges and is only due to the interaction with the geometry of spacetime.

The idea of zero-point oscillations of the field is important. In quantum mechanics, the energy of every independent mode of field oscillation is equal to $E_\omega = (n + \frac{1}{2})\hbar\omega$, with n equal to the number of field quanta and the $\frac{1}{2}\hbar\omega$ being unobservable. The vacuum corresponds to $n = 0$, $E_\omega = \frac{1}{2}\hbar\omega$.

Assume that the classical theory gives the result, that a wave with E, ω is transformed into E', ω', so that $N'_c = E'/\omega'\hbar > N_c = E/\omega\hbar$ (index c for classical). In this case the correspondence principle leads to the quantum result that the vacuum with $N_0 = \frac{1}{2}$ is transformed into a state with $N'_0 = \frac{1}{2}N'_c/N_c$, $E'_0 = \frac{1}{2}(N'_c/N_c)\hbar\omega'$. The $\frac{1}{2}\hbar\omega'$ is unobservable, but the surplus equal to

$$\Delta E = \frac{1}{2}\left(\frac{N'_c}{N_c} - 1\right)\hbar\omega'$$

describes the creation of real, observable particles from the vacuum.

Shortly, having solved the classical field equations and having established definite departures from adiabatic invariance, we obtain quantitative information about the quantum process of particle creation in a vacuum.

Suppose that the metric is given by the expression

$$ds^2 = dt^2 - [1 + h(t)]dx^2 - [1 - h(t)]dy^2 - dz^2 .$$

Let the function $h(t)$ be $h(t) = \delta \cdot f(\omega t) = \delta e^{-\omega^2 t^2}$ or $h(t) = \delta(1 + \omega^2 t^2)^{-1}$, so that $h = 0$ for $t = \pm\infty$; $h \sim \delta$; and moreover $\delta \ll 1$. It then turns out that, to first order with respect to the small quantity δ, the anisotropic stresses

$$T_{11} = -T_{22} = \delta\omega^4\psi(\omega t) , \qquad T_{33} = \epsilon = 0$$

arise; the function ψ also goes to zero for $t = \pm\infty$. To second order in the quantity δ, we find

$$T_{11} = \delta\omega^4\psi(\omega t) + \delta^2\omega^4 g_1(\omega t) , \qquad T_{22} = -\delta\omega^4\psi(\omega t) + \delta^2\omega^4 g_2(\omega t) ,$$

$$T_{33} = \delta^2\omega^4 g_3(\omega t) ; \qquad T_{00} = \epsilon = \delta^2\omega^4 g_0(\omega t) .$$

The functions g_1, g_2, g_3, and g_0 do not vanish as $t \to \infty$. They are all positive and satisfy $g_0 = g_1 + g_2 + g_3$ in accordance with the law $T \equiv T_i^i = 0$ for quanta with zero rest mass.

The first-order terms violate the energy dominance conditions during the period when metric changes are important! They describe vacuum polarization. The energy conservation condition is

$$\frac{d\epsilon}{dt} = h(T_{11} - T_{22}) .$$

The first order in T_{11} and T_{22} gives a second order contribution in ϵ, which is not vanishing at $t = +\infty$ and which describes creation of real particles (see Zel'dovich and Pitayevskij 1971). For a typical power-law dependence of the metric on time, the characteristic time for a change in the metric equals the time t elapsed since the time of the singularity. Consequently, the waves are nonadiabatic, with $\omega < t^{-1}$ and $k < (ct)^{-1}$. Assuming that creation takes place on the average at the rate of one quantum per mode in this region, we find that the order of magnitude of the energy density of created quanta is $\epsilon \approx \hbar\omega k^3 \approx \hbar/(c^3t^4)$.

Note that the quantity G does not enter the result even though the question is that of the creation of particles in a gravitational field! Notice further the strong dependence of ϵ on t. What does it mean? Strictly speaking, we have found in order of magnitude the energy density of particles created during the time between t and $2t$ (or $te^{-1/2}$ and $te^{1/2}$). Here a great difference arises between the problem of collapse (with a singularity in the future) and the cosmological problem (with a singularity in the past).

In the collapse problem, the time period involved is that when the time is negative since the singularity time is set at $t = 0$. At a given instant t_0, particles created long before (in the period earlier than $2t_0$, for example) make a small contribution to ϵ. But the rate of particle creation grows rapidly. At each given instant the particles created during the most recent times play the main role. For example, if we keep in mind that $t < 0$ and $t_0 < 0$, then we realize that the particles created in the interval $2t_0 < t < t_0$ play the main role. The formula $\epsilon = \hbar/(c^3t^4)$ thus holds, although only as an order of magnitude estimate.

Examining the collapse problem further, one can ask when the created particles themselves significantly affect the metric. Of course, we have until now been considering only the propagation of "test" waves (compare with "test" particles) in a specified metric. Thus, in the GTR equations, power-law solutions correspond to behavior such that the components of the curvature tensor are of the order of $(ct)^{-2}$. On the right-hand side of the GTR equation, meanwhile, the quantity $G\epsilon/c^4$ is found. Substituting the expression for ϵ and equating the right and left sides, we obtain the characteristic time that is expressed in terms of G, $\hbar$, and c, and which consequently cannot differ from t_{Pl}.

In the collapse problem, therefore, it is already clear what is new: one must apply quantum-gravitational theory to it only in the very last moment.

Upon the approach to a singularity, new particles—photons, electron-positron pairs, neutrino-antineutrino pairs, and gravitons—are created, thereby manifesting the destruction of the adiabatic invariance. For $t \to 0$, their energy density grows more rapidly than does the energy density of the preexisting matter, the matter that fills the space before the singularity develops and that contracts according to an adiabatic law. Upon the approach to t_{Pl}, the effect of newly created particles becomes predominant

and they influence further changes in the metric even if, prior to t_{P1}, the preexisting matter does not influence the metric and there is effectively a vacuum stage of approach to the singularity (see § 19.3).

Finally, we can in principle advance further, into the interval when $t_{P1} < t < 0$, by solving jointly the equations for the change in the metric and the equations for describing the particle creation.

A totally different situation arises upon an attempt to apply the theory of particle creation to cosmology. Let us begin with a consideration of an instant t_1 at which, we assume, the metric is specified. In a spatially homogeneous problem, for example, what is specified at this instant are the values of the curvature and of the expansion rate in the various directions and the structure constants characterizing the type of space. We shall neglect the energy density and momentum density of the matter present at the instant t_1 in accordance with an assumed vacuum character for the solution.

Thus, over the time from t_1 to $2t_1$, particles with an energy density whose order of magnitude is $\epsilon \sim \epsilon_1 = \hbar/(c^3 t_1^4)$ arise in the vacuum. We emphasize that this formula holds for a very short time in the cosmological problem. At the later instant t_2, with $t_2 > t_1$, the value of ϵ resulting from the most recently created particles satisfies $\epsilon \sim \epsilon_2 = \hbar/(c^3 t_2^4)$. However, the particles created earlier (at $t \sim t_1$) do not disappear; rather, the volume they occupy expands, so their energy density becomes $\epsilon \sim \epsilon'_1 = (t_1/t_2)^{4/3} \times \hbar/(c^3 t_1^4)$. Since $\epsilon'_1 > \epsilon_2$ therefore, in distinction to what is true for the collapse problem, the energy density at a given instant t depends entirely on both the instant when the particle creation turns on and the manner in which the creation turns on.

To summarize then, at least for a certain time (before $t = t_{P1}$ and perhaps even later) in the collapse problem, an analysis is possible of the phenomenon without regard to the limitations of existing quantum-gravitational theory. In cosmology, on the other hand, the Universe "remembers" the initial conditions at every instant. Along with these general considerations, one should also note an important specific fact. In the theory of wave propagation—and as a consequence in the theory of particle creation too—the quite important principle of conformal invariance applies. This principle allows one to go further. In particular, it allows one to explain the qualitative differences between the isotropic singularities in Friedmann and the anisotropic Kasner singularities. Recall that Lifshitz, Khalatnikov, and their co-workers, having posed the question of the character of the general solution of the GTR equations, reached the conclusion that a singularity is in general anisotropic. In the collapse problem, a general solution must be realized. In the cosmological problem, a particular solution could be realized near the singularity.

A change in a metric is termed conformal if it consists of a change in the scale of all lengths and times that can be different for different world points but must be the same for all spatial directions and for time at a given point.

So, for example, one can transform the Minkowski metric into a "conformally flat" metric as follows:

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2 \to ds^2$$
$$= a^2(x', y', z', t')[dt'^2 - dx'^2 - dy'^2 - dz'^2] .$$

We emphasize that upon such a transformation not only the coordinate frame, but also the geometry, changes. In essence, the question is not of a coordinate transformation, but of the establishment of a correspondence between the various four-dimensional manifolds. A "conformally flat" metric has a nonzero curvature tensor $R_{\alpha\beta\gamma\delta}$ that is expressed in terms of derivatives of the function $a(x, y, z, t)$. In a conformally flat metric, the propagation of a wave that moves at the speed of light is easily examined. Thus, a ray which obeys the condition $ds = 0$ corresponds to a solution $\exp(i\boldsymbol{k}\cdot\boldsymbol{x} - i|\boldsymbol{k}|ct)$ in the Minkowski metric. But such a solution holds in a conformally flat metric too. For if $ds = 0$, then $a(x, y, z, t)ds = 0$ also. Now wave propagation in the flat Minkowski metric is not accompanied by particle creation. There is consequently no creation of massless particles in a conformally flat metric either.

The initial stage of a Friedmann model is described by the metric

$$ds^2 = dt^2 - a^2(t)(dx^2 + dy^2 + dz^2) .$$

Such a metric is conformally flat. To see this, we introduce

$$\eta = \int \frac{dt}{a(t)} , \qquad d\eta = dt/a(t) ,$$

and express $a(t)$ as a function of η: $a(t) = b(\eta)$. We obtain finally

$$ds^2 = b^2(\eta)[d\eta^2 - dx^2 - dy^2 - dz^2] ,$$

which is just the required form. On the other hand, one cannot bring the Kasner solution

$$ds^2 = dt^2 - t^{2p_1}dx^2 - t^{2p_2}dy^2 - t^{2p_3}dz^2$$

to such a form, so it is not conformally flat.

In the Friedmann solution, particles with zero rest mass are not created, while particles with nonzero rest mass are not created in any significant number. The dimensional estimates made above with $\epsilon \sim t^{-4}$ in reality apply only to an anisotropic singularity.

One can evidently interpret this result in hydrodynamical terms. One can term particle creation a manifestation of the viscosity of the vacuum: Upon a deformation of the vacuum, heat is released and entropy grows. Now two types of viscosity are known in hydrodynamics: the "first," connected with a shear deformation of a volume element of fluid; and the "second," connected with a change in the fluid density, i.e., with an overall

uniform expansion or contraction of the fluid. It is well known, for example, that an ultrarelativistic gas does not have a second viscosity.

One can also carry this result over to a "vacuum of ultrarelativistic particles," i.e., to the problem of particle creation. In the Kasner solution, a shear deformation occurs and particle creation takes place. In the Friedmann solution the expansion is isotropic, so only the second viscosity could act; it is absent, however, so particle creation does not take place. About particle creation in the Friedmann model see L. Parker (1968, 1969, 1971, 1972, 1973), Grib and Mamaev (1969, 1971), Chernikov and Shavokhina (1973), Mamaev, Masstepanenko, and Starobinsky (1976); about this problem in anisotropic models see Zel'dovich (1970*c*), Zel'dovich and Starobinsky (1971), Hu, Fulling, and Parker (1973), Hu (1974), and Berger (1974).

We discuss only the case $Gm^2/\hbar c \ll 1$; for the case where there are particles with $M \approx (\hbar c/G)^{1/2}$ see Stanukovich (1965, 1966*b*), Markov (1966).

One can advance a hypothesis according to which an isotropic emergence from the singularity occurs in nature simply because particle creation would lead to internal contradictions in the theory otherwise. (See Zel'dovich 1970*c*; Lukash and Starobinsky 1974.) Thus, in the case of collapse, it follows from general principles that particle creation decreases the anisotropy of the contraction. In cosmology one can assume that the effect has the same sign and, that upon the turning on of the particle-creation process at an instant $t_1 \to 0$, the residual anisotropy also tends to zero.

More precisely, one ought to assume that the true solution is locally isotropic. For the Universe as a whole, such a consideration leads to a quasi-isotropic solution, whose properties were described earlier. It was remarked in the same discussion that these properties (particularly the nucleosynthesis) agree well with what is known about the present-day Universe. What remains unknown is the scale and amplitude of deviations of the metric from homogeneity. Definite nontrivial results do exist, however, for example, that vorticity is absent. Let us be more precise. In a quasi-isotropic solution near the singularity, the vorticity turns out to equal zero for all time (for $t > 0$) and for all spatial points if the matter filling the Universe is an ideal fluid with a definite Pascal pressure $P = f(\rho)$, if the matter is viscosity-free and if its entropy is constant.

Thus, deep theoretical considerations can in principle lead to consequences that are important for the early stages. We emphasize, however, that at the present time what we are discussing remains at the level of an hypothesis.

In such a conception as has just been outlined, however, the magnitude of the entropy remains without an explanation. And in a good theory, of course, one would like to have an explanation of all important properties of the Universe. Moreover, the nature of the spectrum of perturbations leading to galaxy formation remains without an explanation.

In conclusion, note that the questions touched upon here in general terms qualitatively and in the language of general physical ideas are considered more quantitatively in the next sections. The authors are not laboring under any delusions, however. Setting forth a mathematical account of the hypotheses neither makes them reliable nor proves that they are in reality valid. This is so especially inasmuch as the deepest questions remain without answers.

24.4 The Creation of Charged Particles in Electrodynamics

Particle creation in gravitational fields is an extraordinarily exotic process from the viewpoint of laboratory physics. There is no hope whatsoever for an experimental study of the process. On the other hand, particle creation in electromagnetic fields is observed experimentally. The theoretical predictions of modern quantum electrodynamics have been verified and confirmed by experiment. One can therefore consider quantum electrodynamics as an "elder brother"—as a fine example of a theory of quantum-gravitational phenomena. Let us therefore recall briefly some ideas and results that are well known in electrodynamics. A significant number of them have analogs in gravitational theory.

The reader should have no illusions about this discussion—the authors certainly have none: what follows will not teach one how to calculate the effects and will not replace the relevant textbooks, such as that of Akhiezer and Berestetsky (1969) and that of Berestetsky, Lifshitz, and Pitaevsky (1968).

The annihilation process $e^+ + e^- \rightarrow 2\gamma$ has been well studied experimentally. According to general principles of quantum theory, the possibility of the reverse process $2\gamma \rightarrow e^+ + e^-$ follows from the existence of the first. In particular, if the kinetic energy of the e^+e^- in the center-of-momentum system is of the order of $m_e c^2$ (requiring relativistic, but not ultrarelativistic, e^+ and e^-), then the cross sections for the direct and the reverse processes are of the same order, namely, of the order of $e^4/(m_e{}^2c^4) = r_0{}^2$, where r_0 is the so-called classical electron radius, $e^2/r_0 = m_e c^2$.

In the language of Feynman diagrams, e^-e^+ pair creation is described in Figure 62. The time evolution of the events in this figure proceeds from

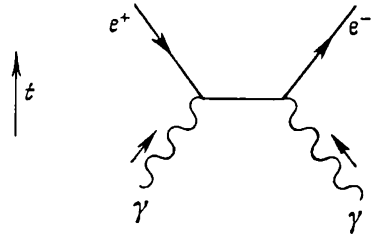


Fig. 62.—Feynman diagram for e^+e^- pair creation

the bottom to the top. The wavy lines represent the photons participating in the reaction. The continuous lines depict the motion of the electron and the positron. One can also describe the positron as an "electron moving in the opposite direction with respect to time."

The dependence of the wave function of the participating particles on time is determined by the factor $e^{-iEt/\hbar}$. A change in the sign of time (or of the direction of the arrow in Figure 62) is equivalent to a change in the sign of energy. According to Dirac, therefore, a positron is a hole in a negative-energy sea of electrons. These considerations are not a proof, but perhaps they can be intuitively reconciled with the idea that one can describe a normal positron, which moves forward in time, by the line of an electron moving backward in time—from the future to the past.

Two photons participate in the pair-creation processes. There are hence two "vertices" in the diagram, two points at which one photon and two electron lines are tied. At each vertex the interaction is proportional to the electron charge e, so that the net effect for the two vertices is a matrix element proportional to e^2 and a cross section or probability for the process proportional to e^4.

Notice that we can obtain Figure 63, the diagram for the process $e^- + \gamma \rightarrow e^- + \gamma$, by "laying Figure 62 on its side." This new diagram represents the Compton scattering of a photon by a free electron. It turns out that diagramatically and mathematically too, the two processes, pair creation and Compton scattering, are closely connected.

Does pair creation also exist in the lowest or first order, i.e., does it occur in accordance with the diagram of Figure 64? A vertex of the type in Figure 64 is possible. But in this case the electromagnetic field, which is depicted by the wavy line, cannot be the field of a free photon, the field of a freely propagating electromagnetic wave. The point is that a free photon is characterized by the relativistic relation $E = pc$ between its energy E and momentum p. Thus, the creation by a free photon of the pair $e^+ + e^-$, for which

$$E_{e^+} + E_{e^-} > (P_{e^+} + P_{e^-})c\,,$$

is incompatible with the conservation laws.

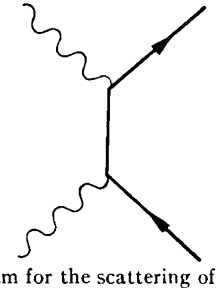


FIG. 63.—Feynman diagram for the scattering of a photon by a free electron

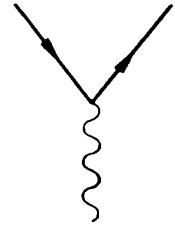


FIG. 64.—Feynman diagram for the creation of an e^+e^- pair by an oscillating electric field.

This does not mean, however, that the process with the vertex as in Figure 64 is entirely impossible. A photon's energy is specified and is everywhere the same if the oscillation frequency of the system emitting the photon is specified. The photon's momentum is specified and corresponds to that in the formula presented above only far from the radiating system in the plane wave approximation. At a distance r from the radiating system the indeterminacy in the momentum is of the order of $\Delta p = \hbar/r$. The electromagnetic field near the radiating system (at a distance of the order of characteristic length $\hbar/[m_e c]$) is therefore capable of creating pairs according to the diagram of Figure 64. Indeed, e^+e^- pairs can be created in this fashion when there is no radiation of real photons (when there are no photons capable of going to infinity) at all!

Such a situation arises in zero-zero transitions. Here the first zero is the spin of an excited system and the second is the spin of the system in the ground state. A spin equal to zero signifies complete spherical symmetry for the system. Classically, one can imagine a spherically symmetric charged cloud undergoing radial oscillations. Outside there is evidently no external radiation—nor are there any variations of the field—because a radiation wave field must be transverse (perpendicular to the direction of propagation). The outside field is completely defined by the total charge of the symmetric cloud. However, inside the cloud there is a variable radial field. Such a field is capable of giving rise to pair creation. The process is observed in experiment in well-known excited states of carbon ($^{12}C^*$) and oxygen ($^{16}O^*$), which transform into the ground state by means of the emission of an e^+e^- pair.

In all of the cases so far described, the energy of the photon or photons, free or not, exceeds the sum of the masses of the e^+ and e^- particles; otherwise the creation would be energetically impossible. But photon energy is connected with frequency by the relation $E = \hbar\omega$. Hence, a field with radiation of sufficiently high frequency is needed in the process above. We must then have

$$\omega > \frac{2m_e c^2}{\hbar} \quad \text{or} \quad \omega_1 + \omega_2 > \frac{2m_e c^2}{\hbar}.$$

The quantum nature of the photons is not important here!

Imagine the specific electromagnetic wave of the type emitted by a laser. On the one hand, one can say that the occupation number N, the number of photons in one mode of the electromagnetic field, is large. On the other hand, one can talk about classical electromagnetic fields for $N \gg 1$. If the probability for a process is computed for one photon (or for two oppositely moving photons), then going now to consider an N-photon state, the probability increases by a factor of N (or N_1N_2 given the collision of two beams with the occupation numbers N_1 and N_2).

One thus obtains an expression for the probability of pair creation in classical electromagnetic fields that depends on the energy density. For the diagram of Figure 64, it is proportional to e^2E^2; and for the diagram of Figure 62, to e^4E^4, where in the latter case we are not distinguishing between the energies in the two oppositely moving beams, whence $E_1 \sim E_2 \sim E$. The field must be of high frequency $\omega > 2m_ec^2/\hbar$ in the Figure 64 case and $\omega > m_ec^2/\hbar$ in the Figure 62 case. In such a theory, a condition on the frequency containing Planck's constant $\hbar$ is a consequence of the application of quantum mechanics to the behavior of electrons and positrons.

Is pair creation possible in a static field, one with $\omega = 0$? Such a process is possible, but it proceeds by a mechanism that is in principle different. Imagine therefore a constant (in space and time) electric field E. Suppose that a pair is created at the point x_0. For this to happen, an energy $2m_ec^2$ is of course necessary. Now the displacement of the positron from x_0 to x_1 is accompanied by a release of energy $e|E|(x_1 - x_0)$. The displacement of the electron from x_0 to x_2 is similarly accompanied by an energy release $-e|E|(x_2 - x_0)$. Thus the necessary expenditures of energy upon the creation of a positron at rest at x_1 and an electron at rest at x_2 is

$$\Delta E_0 = 2m_ec^2 - e|\boldsymbol{E}|(x_1 - x_2) .$$

If $x_1 - x_2$ is great enough, the process is possible without any kind of additional energy source. If $x_1 - x_2 > 2m_ec^2/(e|\boldsymbol{E}|)$, then $\Delta E_0 < 0$.

This means that we have no need of extra energy; of course, an excess energy is present. Therefore the electron and the positron are moving with finite velocities so that the energy of each of them is greater than m_ec^2. The creation of pairs by a static field is thus energetically possible. However, the first requisite condition is that the region of space occupied by the field be sufficiently large that the potential difference is greater than $2m_0c^2/e$, or greater than 10^6 volts numerically. The e^+ and e^- are then created at a definite distance away from each other, and only after the creation do they accelerate. One evidently cannot imagine classically the simultaneous emergence of different charges at different points of space. Not only is the causal connection then destroyed, but so also the charge conservation is violated (though only at each point separately—not in the system as a whole).

Consequently, the use of quantum mechanics to describe the motion of

the electron and the positron is necessary for an understanding of the process. The e^- and e^+ pass through the region of space (for small $x_1 - x_2$) where, by virtue of energy considerations, they cannot be found classically by tunneling through the energy barrier. From these considerations one can determine the probability of pair creation. The main factor is $W \sim \exp(-\beta \Delta x m_e c/\hbar)$, where $\Delta x = 2m_e c^2/eE$ is the width of the barrier, $\hbar/(m_e c)$ is the characteristic wavelength 3.7×10^{-11} cm, and β is a numerical coefficient of order unity. The characteristic field strength for which the creation probability is not excessively small is of order 10^{16} volts cm^{-1}.

Such a process has up until now not been observed. But the development of laser technology allows one to hope confidently for its realization in the next ten or twenty years. (See also Bagrov, Bozrikov, and Gitman 1973.)

Nikishov (1969) and Narozhny and Nikishov (1970) have carried out detailed calculations of the process.

In the immediate neighborhood of an atomic nucleus, the electric field is significantly greater than the characteristic magnitude mentioned above. Near the uranium nucleus it reaches 10^{18} volts cm^{-1}! A nuclear field decreases rapidly with distance, however. The spontaneous creation of e^+e^- pairs in a nuclear field could therefore take place only for hypothetical nuclei with a charge or atomic number greater than 180.

The creation of pairs in the electric field of nuclei is considered in Zel'dovich and Popov's (1971) review; also listed there are citations to original works. What is shown is that one cannot think of the pair creation by a static field as a sum of the processes described by diagrams of the type in Figure 64 (with one photon line), Figure 62 (with two photon lines), and so forth. This is manifested formally by the impossibility of expanding the expression for probability of the process, $W \sim e^{-E_c/E}$, in the form of a power series with respect to E.

Turn now to the limiting case of a low-frequency or weak field. In other words, consider conditions under which one can neglect the creation of real pairs of electrons and positrons. Quantum mechanics then predicts a variety of phenomena—virtual pair creation, for example. One can intuitively imagine that pairs are created but then immediately disappear. No free charged particles are ever present, but there nevertheless appears a nonzero, time-averaged charge density and a nonzero, time-averaged electric current. This phenomenon is customarily called the "polarization of the vacuum." It is useful now to turn to the theoretical bases of this phenomenon.

The theory of elementary charged particles is symmetric with respect to the sign of the charge. The "positive" sign for the positron charge is merely a convention. The prevalence of electrons and protons (at least in our neighborhood) thus reflects the initial conditions of the Universe, and not the local laws of physics.

It follows from this that the vacuum must be electrically neutral. However, the electric and magnetic fields evidently destroy the charge symmetry of the vacuum. (A philological question remains: Should one call a space in which an electromagnetic field exists a vacuum? We shall agree to do so here. In a vacuum which may contain an electromagnetic field, there must not be free charged particles. Consequently there may be photons present. More precisely, therefore, we speak of a "vacuum with respect to charged particles" here. In the main chapters of the book, which are devoted to the theory of gravitation, we mean by "vacuum" a space in which no kind of particle or field whatsoever is present.) In the presence of fields, charges and currents appear. These charges and currents must express themselves through an electromagnetic field. But then the most general types of properties—the vector and tensor character of the quantities under consideration, the invariance of the theory with respect to calibration of the potential, and so forth—strongly limit the form of the formulae connecting the fields and the vacuum currents. Let us now explore this in more detail.

We shall be characterizing the electromagnetic field by means of the four-vector A_μ, where $A_0 = \phi$. Along with A_μ, the quantities $\Box A_\mu$, $\Box^2 A_\mu$, and so on are also vectors. For the vacuum current density, consequently, we can propose the formulae

$$j(\text{vac}) = \alpha_0 A\ , \quad j(\text{vac}) = \alpha_1 \Box A\ , \quad j(\text{vac}) = \alpha_2 \Box^2 A\ .$$

Of these formulae, the expression with α_0 drops out by virtue of the guage invariance. That is, only the electric and magnetic fields themselves have a physical sense, can be measured locally, and are unambiguously defined. But the vector potential A_μ is ambiguously defined, so the expression with α_0, which would yield ambiguous potential values, cannot hold.

The situation regarding the α_1 expression is curious. According to Maxwell's equations

$$\Box A = 4\pi j = 4\pi(j_{(\text{ext})} + j_{(\text{vac})})\ .$$

On the right-hand side stands the total current density, i.e., the sum of the "external" current density of the charged particles $j_{(\text{ext})}$ and the vacuum current density, $j_{(\text{vac})}$. If $j_{(\text{vac})}$ is in turn given through the formula with α_1, then we obtain

$$\Box A = 4\pi j_{(\text{ext})} + 4\pi\alpha_1 \Box A\ ; \quad \Box A = \frac{4\pi j_{(\text{ext})}}{1 - 4\pi\alpha_1} = 4\pi j_{(\text{obs})}\ .$$

All of the formulae of electrodynamics are preserved, but the observable current density $j_{(\text{obs})}$, which is proportional to $j_{(\text{ext})}$ and differs from it by the factor $(1 - 4\pi\alpha_1)^{-1}$, replaces $j_{(\text{ext})}$. The observable charge of each elementary particle differs by the same factor from the "true" or "original" or "bare" charge. A so-called renormalization of the charge thus occurs.

The renormalization coefficient is the same for all charged particles. It is therefore impossible to determine it experimentally. Calculation shows that

$\alpha_1 < 0$ and that $j_{(obs)} < j_{(ext)}$. In other words, the vacuum polarization partially screens the bare charge. But the result of a consistent calculation (Landau and Pomeranchuk 1955) is that the magnitude of the screening diverges logarithmically. In this sense, a logically closed, consistent formulation of quantum electrodynamics does not exist. One can, however, carry out calculations in quantum electrodynamics in such a way that only the observable charge of particles enters into the formulae for all observable quantities (scattering cross sections, energy levels, and so on). It is just the theory so constructed that indeed agrees wonderfully with experiment.

Let us be more precise. There does not exist a theory without a cutoff parameter for small lengths or large momenta. If, however, such a parameter Λ is introduced, to eliminate divergences, then the ratio $j_{(obs)}/j_{(ext)}$ satisfies $j_{(obs)}/j_{(ext)} \sim \{1 - [e_{(obs)}^2/(3\pi)] \ln [\Lambda/(m_e c)]\}^{1/2}$. Errors in the theory, in which $e_{(obs)}$ manifestly plays a role, are then of the order of $m_e c/\Lambda$. A theory with $\Lambda \to \infty$ (a "closed" theory) does not exist, but a theory with $\Lambda \sim m_e c \exp(\hbar c/e^2)$, a theory satisfactory in practice, does indeed exist.

The next formula, that with α_2, describes real observable effects. Now one of the components of the four-vector j is the charge density. The formula with the α_2 term gives a vacuum charge density that differs (as a function of the coordinates) from the external charge density.

Experimentalists refer to this α_2 term as the "vacuum polarization" term. It changes the distribution of the potential around an atomic nucleus, and as a consequence the atomic levels are shifted. Such a shift has indeed been observed. The effect under consideration amounts to about 2.5% of the entire quantum-electrodynamical shift in the hydrogen atom. However, the precision of the computations and of the observations is so high (better than 10^{-7}) that the contribution of the vacuum polarization has been established beyond any doubt. In ions with a large nuclear charge and, in particular, in ions with a muon instead of an electron in an orbital, the vacuum-polarization effect is even larger than for ordinary, neutral atoms and therefore easier to observe.

Only the terms of $j_{(vac)}$ linear with respect to the potential and the field were considered above. In the language of diagrams, the situation described corresponds to that presented in Figure 65. The line of the electromagnetic field is interrupted by the formation of a virtual e^+e^- pair (the current $j_{(vac)}$ arises), after which the pair is transformed back into a field. Thus $j_{(vac)}$ appears on the right-hand side of the equation for A. One can think of the charge renormalization as a consequence of the photon's (or, in a more general formulation, the field's) acting like an e^+e^- pair for part of the time. Inasmuch as the photon (or the field) does not interact with the other, real, charged particles during this time, the effective interaction—and hence the observable charge—is less than it would otherwise be.

In addition to linear effects there are also terms in the charge-density expression proportional to the third power of the electromagnetic field

strength (and even to higher odd powers). Taking into consideration the interaction of this additional contribution to $j_{(\mathrm{vac})}$ with the field, we obtain effects corresponding to a diagram with four photon (field) lines (Fig. 66). Such a diagram describes the scattering of light by light, still unobserved, and also the scattering of light by the electric field of a nucleus. The latter process, predicted by Delbruck, has been successfully observed in the laboratory.

One can interpret these processes as the manifestations of the nonlinearity of electrodynamics. In effect, the terms $\sim(\boldsymbol{E}^2 - \boldsymbol{H}^2)^2$ and $\sim(\boldsymbol{E}\cdot\boldsymbol{H})^2$, in addition to terms of still higher order, appear in the field Lagrangian. Correspondingly, cubic terms appear in Maxwell's equation.

There is a case in electrodynamics when one can consider the field to be classical. Thus, it was already mentioned that the field can act in this fashion for the case of large occupation numbers. It is also not necessary to quantitize the electromagnetic field in the cases of, for example, the Coulomb field of a nucleus or a constant magnetic field. Quantum effects in

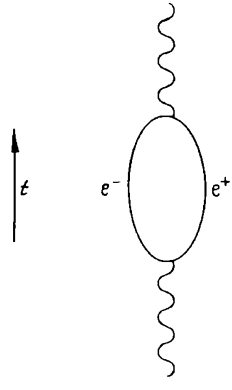


FIG. 65.—Feynman diagram for the formation of a virtual e^+e^- pair

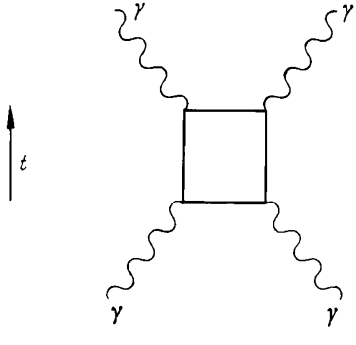


FIG. 66.—Feynman diagram for a nonlinear process in electrodynamics: the scattering of light by light.

these cases are connected only with the application of quantum mechanics to the electrons and the positrons interacting with the fields. If, moreover, the conditions are such that the real creation of pairs does not take place, then one can express these effects in terms of the instantaneous values of the fields and of their derivatives. In this sense one speaks of quantum corrections to Maxwell's equations. The vacuum current is expressed in terms of the field and is placed on the left side of the equations. Only the current arising from real particles remains on the right-hand side:

$$F_{ik;}{}^{k} = -4\pi j_i(\text{total}) \rightarrow F_{ik;}{}^{k} + 4\pi j(\text{vac})_i = -4\pi j(\text{ext})_i .$$

It is evident that such a procedure is possible only so long as the creation of real pairs is not substantial.

The real pairs are not different from the external charges. However, the number of pairs, their motion, and the current to which they give rise are all quantities that depend on the entire previous history of the field. One cannot express them in terms of instantaneous local quantities. For problems involving pair creation, therefore, quantum effects necessarily require changes on the right-hand side of Maxwell's equations.

A consistent description of pair creation includes vacuum polarization as the first stage. This is because only a joint consideration of vacuum polarization and pair creation removes the paradoxes that involve the violation of causality and the local nonconservation of charge. In such a case, it is appropriate to collect all of the effects—both virtual and real—of pair creation in one place, on the right-hand side of the equation. They then constitute the various parts of the vacuum current density.

The vacuum current density at a given instant and at a given point is described by a complicated nonlocal functional of the electromagnetic field in that part of the past, such that disturbances can affect the given space-time point.

The current density j(vac) satisfies the causality principle and the principle of charge conservation. One can thereby assume that Maxwell's equations are not changed. Quantum effects give rise to a nonzero right-hand side (a current density) even in the absence of charged particles in the original state. In particular, such an interpretation, adapted to the case of large effects and real pair creation, will serve us as a model for the examination of quantum phenomena in a gravitational field near the singularity.

The moral that follows from the history of the development of electrodynamics reduces to two theses. The first is that one considers those (and only those) effects and correction terms which are direct, necessary, and inevitable consequences of the unchanged initial assumption that the nature of the interaction of the electromagnetic field with charged particles is as originally described. The second point is that experiment in all its detail confirms the theoretical deductions obtained.

Finally, note that in this way the article by Zel'dovich (1972*c*) in the

Wheeler Festschrift *Magic without Magic* can serve as a continuation of this section. A very readable article by Berestetsky (1976) describes the latest achievements.

24.5 The Mathematical Theory of Particle Creation

This section is oriented to the reader who is also interested in the methodological side of the theory just described. The account will follow the work done by Zel'dovich and Starobinsky (1971). All of the general physical principles were set forth earlier in § 24.3. The reader interested only in the results and in the general ideas of the theory can therefore omit what follows in this section without detriment.

Suppose, to begin, that we have a flat, spatially homogeneous metric of the form

$$ds^2 = dt^2 - a^2(t)dx^2 - b^2(t)dy^2 - c^2(t)dz^2 .$$

In the problem of particle creation, we take a, b, and c to be arbitrary functions of the time t. With this metric we shall consider a massless scalar field ϕ that satisfies the conformally invariant equation

$$(\Box - \tfrac{1}{6}R)\phi = 0 ,$$

an equation which corresponds to the Lagrangian

$$\mathfrak{L} = \tfrac{1}{2}[g^{ik}\phi_{,i}\phi_{,k} + \tfrac{1}{6}R\phi^2] .$$

Here $\phi_{,i} \equiv \partial\phi/\partial x^i$. For information concerning the origin and role of the additional term $R/6$, where R is the scalar curvature, see the studies by Penrose (1964) and by Tagirov and Chernikov (1968). In these works it is shown how this additional term guarantees the conformal invariance of the equation.

Notice that the "simplest" wave equation for a spin-$\frac{1}{2}$ field (the neutrino field) and that for a vector field (the electromagnetic field; hence, Maxwell's equations) are conformally invariant without any additions. An examination of them with regard to particle creation would be more complicated, however, than that of scalar ϕ, but with the same qualitative results (if $R/6$ is put in the ϕ equations).

Let us carry out a transformation of the metric and the field to more convenient variables:

$$g_1 \equiv \frac{a}{(abc)^{1/3}} ; \quad g_2 \equiv \frac{b}{(abc)^{1/3}} ; \quad g_3 \equiv \frac{c}{(abc)^{1/3}} ;$$

$$d\tau \equiv \frac{dt}{(abc)^{1/3}} , \quad \psi \equiv \phi(abc)^{-1/3} .$$

In the terms of the new variables, we have

$$ds^2 = (abc)^{2/3}[d\tau^2 - g_1^2dx^2 - g_2^2dy^2 - g_3^2dz^2] .$$

The equation for ψ then takes the form

$$\frac{\partial^2\psi}{\partial\tau^2} - \frac{1}{g_1^2}\frac{\partial^2\psi}{\partial x^2} - \frac{1}{g_2^2}\frac{\partial^2\psi}{\partial y^2} - \frac{1}{g_3^2}\frac{\partial^2\psi}{\partial z^2} + Q(\tau)\psi = 0 ,$$

where

$$Q(\tau) = \frac{1}{18}\left[(g'_1/g_1 - g'_2/g_2)^2 + (g'_1/g_1 - g'_3/g_3)^2 + (g'_2/g_2 - g'_3/g_3)^2\right] .$$

the notation is that $g'_1 \equiv \partial g_1/\partial\tau$.

Spatial homogeneity permits one to seek solutions having the form

$$\psi = e^{i\mathbf{k}\cdot\mathbf{r}}\chi_k(\tau) .$$

The equation for χ then (we shall henceforth omit the index k) takes the form

$$\chi'' + [k_1^2/g_1^2 + k_2^2/g_2^2 + k_3^2/g_3^2 + Q(\tau)]\chi = 0 .$$

It is convenient to call the quantity $[k_1^2/g_1^2 + k_2^2/g_2^2 + k_3^2/g_3^2]^{1/2}$ the frequency and to denote it by $\omega(\tau)$. The conformal invariance of the theory is used above to decrease from three to two the number of functions on which the solution depends nontrivially. Actually three functions enter into the last equation, but they are connected by the one identity $g_1g_2g_3 = 1$. For $g_1 = g_2 = g_3$ [i.e., $a(t) = b(t) = c(t)$] the metric is conformally flat. The ψ and χ are as in Minkowski space and, of course, no particle creation occurs.

We shall seek a solution of the equations by the Lagrange method. That is, we shall seek a form consisting of solutions valid in the adiabatic approximation multiplied by functions of time. By definition, therefore,

$$\chi(\tau) = \frac{\alpha(\tau)}{(2\omega)^{1/2}}e^{-i\int\omega\,d\tau} + \frac{\beta(\tau)}{(2\omega)^{1/2}}e^{i\int\omega\,d\tau} .$$

We now impose an additional condition due originally to Lagrange:

$$\chi'(\tau) = -i\omega\left[\frac{\alpha(\tau)}{(2\omega)^{1/2}}\exp\left(-i\int\omega d\tau\right) - \frac{\beta(\tau)}{(2\omega)^{1/2}}\exp\left(i\int\omega dt\right)\right] .$$

Without a condition of this type on α and β, the decomposition of χ is not unique.

Comparing this with the complete expression for χ', we obtain an equation connecting α' and β'. From this equation, we find that

$$\alpha' = \frac{1}{2}\left(\frac{\omega'}{\omega} - i\frac{Q}{\omega}\right)e^{2i\int\omega d\tau}\beta - i\frac{Q}{2\omega}\alpha ;$$

$$\beta' = \frac{1}{2}\left(\frac{\omega'}{\omega} + i\frac{Q}{\omega}\right)e^{-2i\int\omega d\tau}\alpha + i\frac{Q}{2\omega}\beta .$$

Consider now the general properties of this system. First, it has the exact integral

$$|\alpha|^2 - |\beta|^2 = \text{const} .$$

Next, after having specified the initial conditions $\alpha \to 1$ and $\beta = 0$ at $t = -\infty$, we find by an iterated solution that

$$\beta = \frac{1}{2} \int_{-\infty}^{t} \left(\frac{\omega'}{\omega} + i \frac{Q}{\omega} \right) e^{-2i \int \omega d\tau'} d\tau$$

for small Q/ω and ω'/ω.

These properties have a clear interpretation. In classical theory, the solutions $\alpha e^{-i\int \omega d\tau + i\boldsymbol{k}\cdot\boldsymbol{r}}$ and $\beta e^{i\int \omega d\tau + i\boldsymbol{k}\cdot\boldsymbol{r}}$ correspond to the propagation of waves in opposite directions. As long as $\alpha = \text{const.}$ and $\beta = 0$, we have a solution with one wave changing adiabatically in a variable metric.

The condition $|\alpha|^2 - |\beta|^2 = |\alpha_{-\infty}|^2$ signifies that two oppositely moving waves are created simultaneously, so that the total momentum of the field does not change. If the initial wave carries a momentum $c\boldsymbol{k}|\alpha_{-\infty}|^2$, then waves subsequently appear that carry momenta $c\boldsymbol{k}(|\alpha|^2 - |\alpha_{-\infty}|^2)$ and $c(-\boldsymbol{k})|\beta|^2$, respectively, but such that the sum of their momenta equals zero. Momentum conservation is a natural consequence of the homogeneity of the space.

Another manifestation of the pairwise wave creation is the appearance of an exponential function of twice the frequency in the differential equations relating α and β. Consequently, the amplitude β depends on the Fourier components of the change in the metric, but with the frequency doubled. In quantum language one can say that portions of energy equal to $2\hbar\omega$ are borrowed from an external gravitational field. This corresponds to the creation of scalar-field quanta in pairs, each with an energy $\hbar\omega$.

If a standing wave of the form $\phi \sim \alpha_0 \sin(\omega t + \delta) \sin(\boldsymbol{k}\cdot\boldsymbol{r} + \kappa)$ were specified at $t = -\infty$, then a standing wave would also be obtained at $t = +\infty$. The amplitude of the latter can be greater or less than that of the former, depending on the initial phase.

A metric perturbation occurs during a definite period of time and is characterized by the form of the functions a, b, and c. In the formulation of the problem, therefore, there is no group of time translations with respect to which the problem is invariant. Assume that $a(t = +\infty) \equiv a(t = -\infty)$, and the same for b and c. Still the energy of the standing wave can increase or decrease in a given metric, so a dependence on the initial phase arises.

The results for a complex traveling wave of the form $\alpha_{-\infty} e^{-i\int \omega d\tau + i\boldsymbol{k}\cdot\boldsymbol{r}}$ evidently do not depend on the initial phase. A change in the energy of a traveling wave under the action of a metric perturbation is always positive and equals the change in the energy of a standing wave averaged with respect to the initial phase δ.

The conformal invariance of the theory is manifested in the appearance in the equations of only the differences in the expansion rates, i.e., the differences in the Hubble parameters, along the various axes.

Up until now the problem examined has been that of a classical field. After a Fourier decomposition, therefore, the equation characterizing the

amplitude of a given spatial mode is the equation of a harmonic oscillator. The spectrum of the harmonic oscillator with constant mass and frequency involves equally spaced levels described by the formula $E = \hbar\omega/2 + n\hbar\omega$, where n is an integer. In complete correspondence with this, for the energy of a wave one can take the characteristic values $E = (n + \frac{1}{2})\hbar\omega$ in correspondence with the number of field quanta n with a given wave vector.

It is well known that, in the case of an oscillator, there are far-reaching coincidences between the results of the quantum and the classical theory. Indeed, calculations done in terms of quantum field theory (too complicated for inclusion in this book) lead to results that coincide with those described above, obtained using the classical theory. This requires, however, a correct choice of the initial amplitude $\alpha_{-\infty}$ to the corresponding energy $\hbar\omega/2$. The zero-point energy, however, is of no importance. The quantity of interest is the energy difference between states with different numbers n.

In the cosmological problem, the energy density and other components of the stress-energy tensor depend on the field under consideration. These quantities are obtained by a summation of the contributions of the separate modes, i.e., by an integration over the space of wave vectors $\boldsymbol{k}$. Such integrals turn out to be divergent. For example, the sum of the zero-point energies $\hbar\omega/2$ for all possible oscillations is infinite.

It therefore becomes necessary to renormalize the theory but in such a way that all of the general properties of the results continue to hold. What has been suggested (Zel'dovich and Starobinsky 1971) is to compare each given wave with momentum $\boldsymbol{k}$ to a wave with larger momentum $\nu\boldsymbol{k}$ and corresponding smaller amplitude $\nu^{-1/2}\alpha_{-\infty}$. By subtracting from the energy of the wave with momentum $\boldsymbol{k}$ the energy of the wave with momentum $\nu\boldsymbol{k}$, one eliminates the zero-point energy. One removes the other divergences in an analogous fashion. In the final results one then takes the limit $\nu \to \infty$. The merit of this method is that solutions of the field equations (both for the wave with momentum $\boldsymbol{k}$ and for the one with momentum $\nu\boldsymbol{k}$) are being considered at all stages of the calculation, which guarantees that conformal invariance and the conservation laws are maintained. However, a wave with momentum $\boldsymbol{k}$ is subject to the nonadiabatic action of the gravitational field, whereas the wave with momentum $\nu\boldsymbol{k}$ behaves strictly adiabatically in the limit $\nu \to \infty$. Therefore, the energy of the new waves created by the gravitational field ($|\beta|^2$ and $|\alpha|^2 - |\alpha_{-\infty}|^2$) remains unchanged and does not undergo a subtraction upon such a renormalization procedure. This method is similar to the Pauli-Villars method, but is different in that large masses are not introduced, so that the conformal invariance of the theory is preserved. In principle, this method permits one to follow all intermediate phases, to calculate the vacuum polarization as well as the real particle production, and to check the conservation laws. It was in this manner that the results outlined in § 24.4 were obtained (see also Parker and Fulling 1974, Fulling and Parker 1974).

24.5 Mathematical Theory of Particle Creation

1. Hawking's Mechanism of Black Hole Particle Creation

In the theory of particle creation, a very recent and very interesting development concerns processes in the vicinity of a black hole. The first relevant papers dealt with rotating black holes described by the Kerr metric (see Vol. 1). Penrose (1969) and Cristodoulou showed that rotational energy could be extracted from these holes. In a thought experiment they considered an unstable particle injected into the ergosphere. When the particle decays, one of the new particles is swallowed by the black hole but the other can escape the ergosphere to be observed. Using local conservation laws, it is possible to arrange it so that the escaping particle has more energy and angular momentum than the original one; in this manner, energy and angular momentum can be extracted from the black hole.

Zel'dovich (1971) and Starobinsky (1973) noted that the same result should follow for a classical wave incident on the black hole. This effect (see also Teukolsky and Press 1974) was called superradiance.

From the analogy between classical effects and quantum vacuum processes, it can be foreseen that a rotating black hole in vacuo (without incoming waves) will radiate spontaneously, losing mass and angular momentum. If a massless field is radiated, the characteristic wavelength is of order of the gravitational radius $r_g = 2GM/c^2$. The time scale for angular momentum loss is of order G^2m^3/c^4h. For a cosmological time $\sim 10^{10}$ yr, this corresponds to $\sim 10^{15}$ g black holes (Starobinsky 1973).

At that time, however, it was thought that a nonrotating black hole does not radiate. But, this belief was soon overthrown. Bekenstein (1972) argued qualitatively that since the area of a black hole behaves like entropy, one could introduce a characteristic temperature of the black hole. Then, in a series of brilliant papers, Hawking (1974, 1975) computed the radiation from a (nonrotating) black hole and showed it to be equilibrium blackbody radiation at a temperature $T(K) \approx 10^{27}\, M^{-1}$ (g).

The basis of the idea is that, as with all radiation, the zero-point oscillations are redshifted by the star that collapsed at a time t in the past to form a black hole. The asymptotic redshift relation for the frequency varies as e^{-ct/r_g}. At every instant of time, outgoing waves of frequency $\sim c/r_g$ are modulated. For such waves, adiabatic invariance is broken. As we have explained above, for classical waves the violation of adiabatic invariance corresponds to the creation of particles. The new result of tremendous importance is that the Hawking mechanism permits (indeed, requires) a black hole to disappear completely. The equation for mass loss is

$$\frac{dm}{dt} = -\frac{\alpha}{m^2}; \qquad m(t) = (m_0^3 - 3\alpha t)^{1/3}. \tag{24.5A.1}$$

The evaporation time is

$$t_0 = \frac{m_0^3}{3\alpha}; \qquad t_0\ (\mathrm{s}) \approx 10^{-28} m_0^3\ (\mathrm{g}) \tag{24.5A.2}$$

For solar mass ($1\ M_\odot = 2 \times 10^{33}$ g) black holes, negligible effects are obtained: $T(\mathrm{K}) \approx 10^{-6}$, $t_0 \approx 10^{44}$ yr. For such black holes, accretion will be overwhelmingly more important.

The most important consequences of Hawking's theory relate to black holes of mass $\leq 10^{15}$ g. It is possible that such black holes were formed during the extremely dense phase early in the cosmological expansion. For that reason such black holes are called Primordial Black Holes (PBH for short). Their formation is discussed next.

B. The Formation and Properties of Primordial Black Holes

We have seen that an attractive hypothesis for the initial singularity is that it was chaotic (i.e., inhomogeneous and anisotropic). An attempt to connect the detailed investigations of small perturbations of a homogeneous (isotropic or anisotropic) model with the general case of strong perturbations meets with immense difficulties. There are two important pieces of information we have to guide us: the absence of strong perturbations on large scales (observations) and the suppression of local anisotropy by particle creation (theory). This does not exclude strong metric perturbations of the quasi-isotropic type on small scales (the definition of small scale will be given later).

But* in the case of strongly perturbed metrics quite new and unexpected nonlinear effects are possible. For instance, portions of plasma could form black holes if their density is sufficiently above the average and/or their expansion velocity is sufficiently below the average, i.e., if the situation in that part of the plasma is like that in a part of a closed Universe. We are dealing in this case with a concave three-dimensional metric (scalar curvature ${}^{(3)}R > 0$).

In order to form a black hole inside a radiation-dominated plasma with $P = \kappa\epsilon$ (for example, $P \approx \epsilon/3$), the concave region must be a rather large part of the closed world and must be surrounded by a region of flat metric and plasma. The spatial dimension l must be of the order of the curvature radius, $l\sqrt{{}^{(3)}R} \sim 1$. This is just the condition for a perturbation of the metric to be of order unity.

Without pressure, a part of a closed Universe behaves in the same way independent of where it is embedded. Expansion comes to a stop at a definite minimum density; thereafter compression begins and a black hole is formed. The plasma—which, unlike dust, has pressure—changes this situation. The density and pressure in the flat part are smaller than in the concave part with the result that a rarefaction wave propagates into the concave part. It depends on the dimensionless values of $\kappa = P/\epsilon$ and $l\sqrt{{}^{(3)}R}$ whether the perturbation is dissipated or a black hole is formed.

The hypothesis of formation of black holes in the radiation-dominated

* The remainder of the English translation of chapter 24 was edited by Douglas Macdonald, of Caltech.

prestellar phase of the evolution of the Universe was formulated by Zel'dovich and Novikov (1966, 1968).

The relation between the plasma density and the radius and mass of the black hole when the gravitational collapse begins is given by

$$l \sim ct\,, \quad M \sim \rho l^3\,, \quad \rho \sim \frac{c^2}{Gl^2} \sim \frac{1}{Gt^2} \sim \frac{c^6}{G^3M^2} \sim 10^{16}\left(\frac{M}{M_\odot}\right)^{-2} \sim \frac{10^6}{t^2}\,,$$

where t is measured in seconds. Primordial black holes with masses smaller than the Sun's mass are formed at $t < 10^{-5}$ s, $z \sim 10^{12}$ in the hadron era; to obtain $M < 10^{15}$ g, one needs $t < 10^{-24}$ s, $z \sim 10^{22}$.

In 1966 it was believed that copious formation of such black holes could be dismissed on observational evidence. The point is that during subsequent expansion the black holes behave like massive nonrelativistic particles: they do not lose mass. The ambient plasma, composed of relativistic particles, loses its energy by redshift or (it is the same effect, see Volume 1) as a result of the work done by pressure during expansion.

If one knows that now the density of black holes is less than $\rho_{\mathrm{PBH}} < 10^{-29}$ g cm^{-3} ($z = 0$), then $\rho_{\mathrm{PBH}}/\rho_\gamma < 10^5$ ($z = 0$), and thus

$$\frac{\rho_{\mathrm{PBH}}}{\rho_\gamma} < 10^{-7} \quad (z = 10^{12})\,,$$

$$\frac{\rho_{\mathrm{PBH}}}{\rho_\gamma} < 10^{-19} \quad (z = 10^{22})\,,$$

$$\frac{\rho_{\mathrm{PBH}}}{\rho_\gamma} < 10^{-30} \quad (z = 10^{32},\ t = t_{\mathrm{Pl}} = 10^{-43}\ \mathrm{s})\,.$$

The conclusion was drawn from this that strong perturbations must be absent even on small scales: assume $\rho_{\mathrm{PBH}}/\rho_\gamma = \exp\left[-(\kappa_{cr}^2/\overline{\kappa^2})\right]$, where κ is the dimensionless perturbation of the metric, κ_{cr} the value of κ needed to form a black hole ($\kappa_{cr} \sim 0.5$), and $\overline{\kappa^2}$ the average perturbation. The values of $\sqrt{\overline{\kappa^2}}$ corresponding to the upper limits given above are $z = 10^{-7} \rightarrow \sqrt{\overline{\kappa^2}} < 0.12$; $z = 10^{-19} \rightarrow \sqrt{\overline{\kappa^2}} < 0.07$; $z = 10^{-30} \rightarrow \sqrt{\overline{\kappa^2}} < 0.05$. There was thus an impression of general smoothness of the Universe on all scales including the "astronomical" one with $(\overline{\kappa^2})^{1/2} \sim 10^{-2}$ to 10^{-3}.

Hawking's discovery of evaporation of PBH brought new dimensions to the discussion (see Carr 1976).

Nothing is changed for PBH with $M \gtrsim 10^{17}$ g, their evaporation being slow and giving low energy photons.

But the birth of 10^{15}–10^{16} g PBH would manifest itself in strong specific γ-ray emission. A search is needed of such objects with explosive output of the order of 10^{23}–10^{24} ergs in one second in the form of 100 MeV γ-rays. No such object has yet been found. Moreover, the measurements of the γ-ray

background give much more stringent limits for the number of such PBH than the limits due solely to their gravitation.

The PBH with masses in the interval $10^9\,\mathrm{g} < M < 10^{15}\,\mathrm{g}$ evaporate after the annihilation of positrons and electrons. Their energy output will thus disturb the equilibrium Planckian spectrum of the blackbody relic radiation (RR). Therefore one can be sure that formation of PBH in this mass interval was very limited.

To a cosmologist the most important new idea is the possibility of copious formation of PBH with $M < 10^9\,\mathrm{g}$, evaporating early enough. In the simple case of a radiation-dominated plasma with equation of state $P = \epsilon/3$, $M < 10^9$ corresponds to $t < 10^{-30}\,\mathrm{s}$, $z > 10^{25}$. Therefore strong perturbations of metrics are not excluded on the scale which corresponds now to $l = ctz = 3 \times 10^5\,\mathrm{cm}$ and smaller scale.

Hypotheses with copious early PBH formation ($10^{-43} < t < 10^{-30}\,\mathrm{s}$) can provide new ideas about the baryon excess and entropy of the Universe. This is briefly discussed below.

There are discussions on PBH formation in the case of a soft equation of state, $P \ll \epsilon$ (Carr 1975). Without going into details, one can summarize by saying that these discussions give arguments against such an initial assumption.

We omit here also the discussion of accretion of plasma on an already formed PBH. In 1966 we could not answer the question of whether the accretion would lead to a modest increase of the initial mass or to a growth of mass proportional to the time. However, Hawking and Carr (1974) have shown that a self-similar solution with $M \sim t$ is nonexistent. We hope to confirm the most probable guess of limited mass increase by direct nonstationary numerical calculations.

24.6 Superspace and Mini-Superspace

The quantization of the evolution of the Universe as a whole is the ultimate goal of the theory of superspace first developed in 1960 by Wheeler and his co-workers (for a review see Misner, Thorne, and Wheeler 1973).

To solve this problem, Wheeler used the formulation of quantum theory proposed by Feynman in 1948, a formulation in which the connection with classical theory is clearly manifest. To understand this, recall first that one can formulate classical theory (mechanics, electrodynamics, and even GTR) in the form of a "principle of least action."[2] According to this principle, for given initial and final states of the system under consideration, that motion is realized for which the integral $\int \mathfrak{L} dt$ is extremized. Quantum theory, on the other hand, is formulated in probabilistic terms. Other motions are realizable in quantum theory which are forbidden in the classical theory, e.g., tunneling through a potential barrier.

2. Concerning Hamiltonian formalism, see Ryan (1972).

Feynman suggested that one must consider every single trajectory leading from a chosen initial state to a chosen final state in figuring the transition probability between them. Note that the phrase "every single trajectory" includes, for example, zigzag motion with variable speed in the problem of the motion of a free particle on which no force at all acts. For each trajectory, the integral $\int \mathfrak{L} dt$ is calculated. The probability of the system evolving from a given initial state x_1 to a given final state x_2 is then given by the quantity $|G(x_1, x_2)|^2$, where G is given by a summation taken over all trajectories:

$$G(x_1, x_2) = \sum \exp\left(\frac{i}{\hbar} \int_{x_1}^{x_2} \mathfrak{L} dt\right).$$

Feynman showed in addition that such a formulation is equivalent to the traditional Schrödinger or Heisenberg formulation of quantum mechanics. It is clear that those trajectories for which the phase factor $\hbar^{-1} \int \mathfrak{L} dt$ is almost stationary, differing by less than π over a number of neighboring trajectories, make the largest contribution to the sum. In this case the exponentials are simply added. Conversely, in regions of the set of trajectories where $\hbar^{-1} \int \mathfrak{L} dt$ fluctuates rapidly between positive and negative values, the corresponding exponentials in the sum cancel one another. It is therefore evident that those trajectories that are close to the extremal or classical trajectory give the greatest contribution to the transition amplitude $G(x_1, x_2)$.

One can also apply Feynman's principle to a more complicated system. A complete characterization of a complicated system requires the specification of a large number of parameters, and consequently it is necessary to consider a "trajectory" in a space with a large number of dimensions.

Wheeler boldly generalized this idea to the Universe as a whole—to a four-dimensional manifold, together with the matter contained in this manifold. The resulting space, whose points represent all possible states of the Universe, is called "superspace." A line—a trajectory in this space—represents one trajectory of the evolution of the Universe, a trajectory that generally does not satisfy the GTR equations. However, a calculation of the action integral for all possible trajectories followed by selection of the particular group of trajectories giving the main contribution to $G(x_1, x_2)$ leads, with the correct choice of $\mathfrak{L}$, to the trajectories corresponding to those of GTR. Simultaneously, one can in principle also determine the degree of quantum deviations from the classical trajectory.

Actually carrying out such a program for the general case is a task fraught with enormous difficulties which have not been overcome at the present time. For example, one particular evolution can in general be parametrized in various ways, so one can select a sequence of three-dimensional spatial slices in various ways. But it is not clear how to guard against multiple counting of a single evolutionary track due to the arbitrary coor-

dinate freedom. It is also not clear what trajectory to use for an open, infinite Universe.

Misner posed a more limited but still relevant problem. He considered the evolution of a closed, homogeneous Bianchi type IX universe. The analysis assumes that the Universe remains homogeneous and of the same Bianchi type at all times. The instantaneous state is then described by the three parameters a, b, and c. In this approximation, consequently, the infinite dimensional superspace reduces to a space of three dimensions, Misner's "mini-superspace."

Now one can formulate the quantum theory in this case with the help of the Feynman trajectories in the mini-superspace; but one can also write the Schrödinger equation for a wave function $\psi(a, b, c; t)$. The square of the modulus of this wave function then gives the probability distribution of the model parameters a, b, and c at a given instant t. Now the kinetic energy, which depends upon the momenta P_a, P_b, and P_c or on the corresponding velocities $a' \equiv da/dt$, b', and c', enters into the Schrödinger equation and the Lagrangian. Terms arising from the spatial curvature $R_{\alpha\beta}$, which involve quantities of the type $a^2/(bc)^2$, also enter; they play the role of a potential energy. But near the zeros of a, b, and c, i.e., near the singularity, the potential energy becomes infinite. The peculiar feature of the quantum problem is thus that a transition from cosmological collapse to cosmological expansion is in principle possible. Such a transition would be similar to what occurs when a charged particle is scattered by a Coulomb center. This analogy is not exact, however, for the scattering of a particle is possible in classical theory too; the particle can fly past a Coulomb center. In the classical GTR problem, however, the collapse of a homogeneous Universe inevitably leads to a state of infinite density, where the volume V of the Universe, proportional to abc, goes to zero.

How does the quantum theory of mini-superspace help in this case? To see how, we shall give an estimate of the quantum number of that quantum state to which the average motion of the Universe corresponds at the present time. To begin, a rough estimate gives[3]

$$\epsilon \approx \frac{c_\gamma^2 V}{G}\left(\frac{a'}{a}\right)^2 \approx \frac{c_\gamma^2 (a')^2 a}{G}$$

for the kinetic energy, where we take $a \approx b \approx c$. The momentum p_a corresponding to the coordinate a satisfies $p_a = \partial\epsilon/\partial a' = c_\gamma^2 a' a/G$. In order of magnitude,

$$a p_a = n\hbar\,, \qquad n = \frac{c_\gamma^2 a' a^2}{G\hbar} = \left(\frac{t_0}{t_{\mathrm{Pl}}}\right)^2 = 10^{120}\,,$$

where t_0, the age of the Universe, is 10^{17} s and where $t_{\mathrm{Pl}} = (G\hbar/c^5)^{1/2} = 10^{-43}$ s. The quantum corrections to the motion of the Universe as a whole at

3. Here c_γ is the speed of light and should not be confused with c, one of the parameters of the model. As an estimate of the parameters we take $a' = at_0^{-1}$ and $a = c_\gamma t_0$, with t_0 the age of the Universe.

the present time are of the order of $n^{-1} \sim 10^{120}$, so they are inconceivably small. Now the number n is an adiabatic invariant. Actually, for both the Kasner solution and the Lifshitz, Khalatnikov, and Belinsky solution, we have $V \sim a^3 \sim t$. In agreement with the above remark, therefore, $a^2a' =$ const. holds for these models in the roughest approximation, where again we assume that $a \approx b \approx c$.

To determine the characteristic parameters of the state of greatest contraction in such a theory is much more difficult. The simplest estimate involves assuming that $a \sim b \sim c \sim l_{\mathrm{Pl}} \sim (G\hbar/c^3)^{1/2} \sim 10^{-33}$ cm, which gives

$$\frac{t_1}{t_0} = \left(\frac{l_{\mathrm{Pl}}}{ct_0}\right)^3 = \left(\frac{t_{\mathrm{Pl}}}{t_0}\right)^3, \qquad t_1 = t_{\mathrm{Pl}}\left(\frac{t_{\mathrm{Pl}}}{t_0}\right)^2 = 10^{-163}\ \mathrm{s}$$

for the time t_1 when the contraction stops.

It is possible for the parameters to depend on the details of the anisotropy of the contraction, i.e., on the differences between a, b, and c. This is in contrast to our previous approximation, where we took $a \sim b \sim c$ by averaging over cycles and eras in the Lifshitz, Khalatnikov, and Belinsky solution. Even in the anisotropic case, though, $t_1 \ll t_{\mathrm{Pl}}$ is obtained.

Such a result is not surprising since the Universe is a huge—indeed the largest—object. It is then natural that for the Universe as a whole quantum effects arise closer to the singularity than for its parts. Therefore Misner's theory of mini-superspace, which is based on the GTR equations, can hardly be literally correct. Nevertheless, this approach will probably be revived in the future as more research is done on local quantum effects.

24.7 The Hypothesis of Baryon Nonconservation and the Charge Asymmetry of the Elementary Particles

A radical approach to the puzzle of the charge asymmetry of the Universe consists of a repudiation of the notion of baryon-charge conservation. Simultaneously, with the adoption of this approach, one of the most important motivations behind the search for a theory describing transitions through a singularity loses appeal, i.e., the need for the continuity of baryon world lines. We shall consider here studies in which attempts are made to develop the hypothesis of baryon nonconservation.

In the first attempt by Sakharov (1967*a*) it is assumed that the elementary-particle process $p \rightleftharpoons 2\mu^+ + \mu^-$ is possible. In the second, by Kuz'min (1970), the process $\kappa^0 \to N$ (or $\bar{N}$) $+ l + \bar{l}$ is introduced, here $\kappa^0 =$ neutral fermion, $N =$ baryon, $l =$ lepton. In the most recent attempt, Pati and Salam (1973) are developing these ideas in connection with quark theory.[4] Cosmological implications of Salam and Pati are developed by Okun' and Zel'dovich (1976).

4. A brand new approach to baryon nonconservation by t'Hooft (1976) has not yet been developed in the cosmological context.

The ultimate goal of these works is to obtain the present-day charge asymmetric state under the assumption of an initially charge-symmetric state for the plasma. We emphasize that in order to obtain this result it is necessary, but not sufficient, to introduce baryon nonconservation.

According to the theory of elementary particles, one can transfer a particle from the right-hand side of a reaction formula to the left-hand side if one also changes the particle to its antiparticle. Consequently, on the basis of the baryon-nonconservation hypothesis, processes of antibaryon creation and annihilation take place along with processes of baryon creation and annihilation. Before 1964, moreover, it was believed dogmatically that all probabilities and cross sections are invariant under particle-antiparticle interchange. In a theory embodying this belief, charge asymmetry evidently does not arise even if the assumption is made that baryon charge is not conserved.

At present, however, we know on the basis of precise laboratory experiments that there is not an exact symmetry between the properties of particles and those of antiparticles. According to traditional ideas, the masses of particles and antiparticles are exactly equal while the electric charges of particles and antiparticles are equal in magnitude and opposite in sign. Moreover, the total decay probability of unstable particles is equal to that of unstable antiparticles. In several decay modes, however, a violation of charge symmetry is manifested. For example, it has been reliably established by experiments that the decay probability for the long-lived neutral kaon K_L to decay by the process $K_L \to \pi^+ + e^- + \bar{\nu}$ is greater than that for the antiparticle to decay by the process $\bar{K}_L \to \pi^- + e^+ + \nu$. In principle, consequently, one can imagine that some hypothetical W^+ and W^- bosons, for which several decay channels exist, have different probabilities for decay upon a change in the baryon charge. Suppose that the probability for the process $W^+ \rightleftarrows p + \nu$ is greater than the probability for the process $W^- \rightleftarrows \bar{p} + \bar{\nu}$.

It is guessed that after cooling and annihilation of $p + \bar{p} \to \pi, \gamma, \ldots$, the excess protons survive and give the observed ratio of $p:\gamma$.

This primitive theory is contradictory to the known stability of the proton and other "stable" nuclei. Obviously, if there are at least two channels for the W^+:

$$W^+ \to p + \nu\,, \qquad W^+ \to e^+ + \nu\,,$$

then there should also be a process through virtual W^+ formation:

$$p \to W^+ + \bar{\nu} \to e^+ + \nu + \bar{\nu}\,.$$

Salam and Pati use the quark theory to reconcile the apparent stability of matter at low temperature with strong enough baryon nonconservation at high temperatures. They guess that quarks (with integer electric charge, of course) are unstable against leptonic decay $q \to$ leptons. The charac-

teristic decay constant is first order in a small coupling constant which corresponds to a lifetime of a free quark of the order of 10^{-10} s. But the nuclear particles are baryons made up of three quarks. The decay of one single bound quark is energetically impossible: the mass of two quarks left after this decay is greater than the mass of the initial neutron or proton. Therefore neutrons and protons are unstable only against the simultaneous decay of all three quarks. The probability of this process is proportional to the third power of the small coupling constant which entered to the first power in the free quark's decay. This makes it possible to reconcile the assumed quark's decay with practical stability of nuclear particles. Here one should remember new trends in quark theory which deny the existence of free quarks (so called "confinement" theory).

Cosmology also gives arguments against the hypothesis of free unstable quarks. At high temperatures the reverse reaction leptons $\rightarrow q$ as well as induced reversible reactions $q +$ leptons $\rightleftharpoons$ leptons must be present.

All reactions, with quarks and antiquarks together, must establish thermodynamic equilibrium. In equilibrium, in the absence of baryon conservation, the baryonic charge density must be zero. Instead of leading to an excess of baryons out of a symmetric Universe, we thus fear that the hypothesis of Sakharov, Pati, and Salam will lead to the destruction of baryons even if initial baryonic charge is given (Okun' and Zel'dovich 1976).

New vistas were opened in baryon nonconservation and Universe asymmetry after Hawking's discovery of black hole evaporation. The question arises of whether we must count the baryons captured by a black hole. Bekenstein (1972) argued that the vector meson field outside the black hole vanishes, and thus, unlike the nonvanishing static electric field which keeps track of electric charge, the vector meson field cannot be used to count baryons. Therefore Bekenstein calls his paper "Nonconservation of Baryons in Black Holes." A black hole appears to be a collective graveyard of unknown baryons. But after introducing Hawking's process of evaporation we lose even the graveyard. One can imagine that matter captured by the black hole is put into a separate closed Universe, disconnected from our space after the end of evaporation, when the mass seen by an outside observer is zero. All the rest mass of baryons captured goes into the energy of the neutral "vapor." In the first approximation, the vapor consists of photons, gravitons, neutrinos, and antineutrinos. At lower masses and higher temperatures electrons and positrons, mesons, and also protons and antiprotons are included.

Various cosmological implications are possible, which are not exhausted as of the time of publication of this book:

1) In a cold baryonic universe with large perturbations on small scale, the formation of PBH and their evaporation could be the source of heat and entropy.

2) In a symmetric hot universe (also with large perturbations of small scale) the formation of PBH and their evaporation can lead to charge asymmetry if CP-noninvariant interactions are taken into account (Hawking 1974). This idea was developed by Zel'dovich (1976): by introducing a heavy neutral meson with CP-noninvariant decay among normal and hypothetical heavy baryons and taking into account the accretion of the decay products, one can obtain approximately the observed asymmetry. But it is ironic that after the experimental discovery of CP violation, the very idea of initial charge symmetry of the Universe has lost much of its attractiveness

3) In principle one can imagine baryon nonconservation in two-particle or one-particle decay: $N + N \rightarrow$ leptons or $p \rightarrow e^+ +$ leptons. The baryons can be considered to have gone from our space into a separate closed space in one quantum jump. A discussion of this problem with an attempt to evaluate the probabilities involved is given by Zel'dovich (1976).

24.8 The Cold Universe and the Perturbation Spectrum

We shall now consider the hypothesis that there was cold matter consisting of baryons alone (without an admixture of antibaryons) near the singularity. This hypothesis is opposite to the hypotheses of a charge-symmetric Universe set forth above.

The appearance of opposing hypotheses in one book reflects the difficulty in determining the state of the Universe and its matter near the singularity. A consistent logical and mathematical development of each hypothesis and its agreement with the observations is of course a natural method for elucidating the truth. The revival of the cold-Universe hypothesis, however, is only partly connected with difficulties in the hot models. The hypothesis of a cold Universe fits in a natural way with the assumption of isotropic (Friedmann) character for the expansion near the singularity. In particular, the isotropy signifies a macroscopic ordering of the initial state: the low initial entropy meanwhile corresponds to ordering small amplitude perturbations on a smaller but still macroscopic level.

At the present time, the Universe is hot; its heat content and entropy are large. An inherent part of the hypothesis of a cold beginning for the evolution of the Universe is therefore an explanation of when and how the present-day, large entropy arises: when and how the Universe becomes hot.

It is assumed here that this heating happens at a time t' satisfying $10^{-43}\ \mathrm{s} = t_{\mathrm{Pl}} \ll t' \ll 10^{-6}\ \mathrm{s}$ as a result of the transformation into heat of the energy of small perturbations. The following calculations show that the perturbations must be of the same amplitude as required for the formation of galaxy clusters, but otherwise of very small wavelength. The assumption of a unique perturbation spectrum over a very broad region of wavelengths thus leads to a reasonable explanation of the magnitude of the entropy.

Such a theory (or hypothesis) is especially attractive because it allows for a natural explanation of the very distinctive matter content of the Universe during the hadron era.

From observations and thermodynamics we know that at the present time there are no regions with an antibaryon excess (Steigman 1976). This means that at an early stage in the hadron era there were plenty of baryons and antibaryons but everywhere a slight excess of baryons. This is just the situation anticipated from a hypothesis of initially cold baryons which is developed in this section.

If one assumes in addition that the lepton charge of the original cold matter is of the same order as the baryon charge, or equal to zero, then almost equal numbers of neutrinos and antineutrinos are obtained after the heating. Research on nucleosynthesis indeed leads to the conclusion that a matter content with equal numbers of neutrinos and antineutrinos is most probable.

In order for all aspects of the theory of nucleosynthesis to remain valid, though, it is necessary that the heating of the matter and the transition from a cold model to a hot model take place sufficiently early: the temperature must be greater than 10 MeV not later than $t = 0.01$ s after the singularity. The establishment of an equilibrium, Planck spectrum for the radiation is then guaranteed. Note that the cold-Universe hypothesis leads quite naturally to these results.

What can one assume as the source of the heating? In this section we shall develop an hypothesis that is based on an extrapolation of density and metric perturbations from scales characteristic of clusters of galaxies to the smallest scale—that equal to the average distance between neighboring baryons. These perturbations are then superposed on a homogeneous and isotropic Friedmann solution. It is then assumed that the original cold baryon gas obeys the well-known equation of state (Zel'dovich 1961) for maximally stiff matter:

$$\epsilon = P = n^2/m^2 ,$$

where n is the baryon number density and m is a quantity with the dimensions of mass and of the order of a proton mass. Here, and in the following, the units are chosen so that $\hbar = c = 1$.

The long-wavelength part of the perturbation spectrum is responsible for the formation of galaxy clusters. This means that the initial perturbation of the metric and of the geometry on the smallest scale must be of the order of $b \equiv 10^{-4}$, which corresponds to a density perturbation that achieves the amplitude $b = 10^{-4}$ at the time when the wavelength λ becomes equal to the particle horizon size ct. Now during the period when $\lambda > ct$, a metric perturbation is constant, a result which follows from the absence of interaction between the parts of the Universe located at a distance λ from each other. Meanwhile, the quantity $\delta\rho/\rho$ grows during this period, with the

law for the growth dependent on equation of state. We have $\delta\rho/\rho \sim t^{4/3}$ for stiff matter with $P = \epsilon$, and $\delta\rho/\rho \sim t$ for matter with $P = \epsilon/3$. It is therefore convenient to characterize the initial spectrum by the metric perturbations themselves. There is no need to be more precise about the particular moment of time to which these perturbations refer as long as $\lambda > ct$.

It has been proposed that the initial perturbations of the metric do not depend on scale.[5] But it is necessary to emphasize that "independence" itself signifies a particular type of function. Note too that at the present time this assertion is not based on any fundamental theory. Observations give only indirect hints in support of the hypothesis that b is constant over the limited mass interval from $10^{12}\,M_\odot$ to $10^{23}\,M_\odot$, or what is the same, over the wavelength interval $10^{23}n^{-1/3} < \lambda < 10^{26}n^{-1/3}$. Meanwhile, it is necessary to extrapolate the perturbation spectrum for the hypothesis proposed to a mass equal to the mass of a single baryon, namely, $10^{-57}\,M_\odot$.

Having enumerated all of the assumptions, we now proceed to carry out the calculations. We shall denote by g the dimensionless quantity $Gm^2 = 10^{-38}$. (In conventional units, $g = Gm^2/[\hbar c]$.)

The unperturbed expansion dynamics of the Universe is described by the formulae

$$\rho = n^2m^{-2} = G^{-1}t^{-2}, \qquad n = mG^{-1/2}t^{-1}.$$

Density fluctuations are expressed in terms of the metric fluctuations b (compare with the expressions of § 11.3) as

$$\delta\rho = bG^{-1}\lambda^{-2}.$$

From these formulae, we obtain

$$\delta\rho = bG^{-4/3}m^{2/3}t^{-2/3},$$

which holds for the shortest wavelengths and for the period when $\lambda_{\min} > t$, and consequently when $b = \text{const}$. We thus have $\delta\rho/\rho \sim t^{4/3}$ as expected.

Let us now find the instant t_1 when the perturbations transform into acoustic waves. We have the condition

$$\lambda_{\min} = n^{-1/3} = t_1,$$

which, together with the expression for n, gives

$$t_1 = G^{1/4}m^{-1/2}, \qquad \rho_1 = G^{-3/2}m, \qquad \delta\rho_1 = bG^{-3/2}m, \qquad \delta\rho_1/\rho_1 = b.$$

The last expression could have been written immediately, for it is well known that the dimensionless amplitude of a density perturbation is just equal to the amplitude of the metric perturbation at the instant of transition from self-similar to acoustic perturbations.

5. What is meant is that the dimensionless quantity $h(M)$ does not depend on the scale; the Fourier components h_k do, however. The connection between them is $h = (h_k{}^2 k^3)^{1/2}$ (see chap. 12).

We shall now compute the acoustic energy and then the temperature and the entropy that result from the decay of the acoustic oscillations into other forms of energy. Thus,

$$\epsilon_{\text{acoustic}} = (\delta\rho)^2/\rho = b^2G^{-3/2}m = T^4;$$

$$T = b^{1/2}G^{-3/8}m^{1/4}, \quad T^3 = b^{3/2}G^{-9/8}m^{3/4}.$$

The dimensionless entropy is then

$$S = T^3/n = b^{3/2}G^{-3/8}m^{-3/4} = b^{3/2}g^{-3/8} = 10^{14.25}b^{3/2}.$$

Hence, in order to get the observed value of S, which, to obtain round numbers, we take to be equal to $10^{8.25}$, it is necessary that one choose $b = 10^{-4}$. This is just what was asserted at the beginning of the section.

Consider now the motivation behind the calculation performed. The most important question is whether we have the right to limit ourselves to wavelengths $\lambda \geq \lambda_{\min} = n^{-1/3}$. The answer to this question is related to definite ideas concerning the nature of maximally stiff matter. See Kobzarev and Okun' (1962).

A static vector field, with baryons as its source, makes the main contribution to the energy density of such matter. The mesons which mediate this field are called vectons (Okun' and Kobzarev 1961), and we call the field a vecton field for short. This static field is similar to a nuclear Coulomb field in that the static field does not fluctuate and is not quantized, but the two fields are different in that the range of the vecton field is finite. A vecton field moreover has transverse degrees of freedom, such as is the case for electromagnetic waves, though an excitation of them would signify an initial temperature differing from zero. But we wish to consider only the static field. It depends unambiguously on the distribution of baryons in space. Therefore, the baryons, together with their field, form a system in which the number of degrees of freedom equals the number of baryons. And to this number there also corresponds (as in the theory of the specific heat of a solid body) a choice of the minimum wavelength.[6] In particular, long-wave perturbations of the baryon distribution correspond to density fluctuations. Perturbations with $\lambda = \lambda_{\min} = n^{-1/3}$ represent the short-wave edge of the spectrum whose long-wave part represents perturbations that give rise to galaxy clusters.

A formulation involving metric fluctuations $h_1 = b = 10^{-4} = \text{const.}$ that are independent of scale is easily and conveniently remembered. It is necessary to emphasize, however, that one must here make an arbitrary assumption justified only by a coincidence involving the entropy now to be described.

6. The condition determining $\lambda_{\min}$ is that $V\int_0^{k_{\max}} k^2 dk$, which gives the number of waves in the volume V, equals nV, the number of baryons in this volume, where $k_{\max} = 2\pi/\lambda_{\min}$.

One can formulate the original hypothesis as an assumption concerning the degree of dependence of the density fluctuations on the mass

$$\frac{\delta\rho}{\rho} = bN^{-2/3}g^{1/3}\left(\frac{t}{t_{\mathrm{Pl}}}\right)^{4/3},$$

for $\lambda > \lambda_{\min}$ or $N > 1$ and when $t < t_1$ (so that the equation of state for stiff matter holds). As was already remarked, moreover, the fluctuations we need are less than those which would be obtained according to a naive approach and satisfying $\delta\rho/\rho = \delta N/N = N^{-1/2}$.

Now one can express the instant t_1 as

$$t_1 = G^{1/2}g^{-1/4} = 10^{9.5}G^{1/2}.$$

Here $G^{1/2}$ is the Planck time t_{Pl}, equal to 0.5×10^{-43} s. It is apparent from this expression for t_1 that important processes occur significantly later than t_{Pl}. In particular, some occur for $t_1 = 10^{-34}$ s, by which time no effects connected with the quantization of spacetime itself or with corrections to GTR are manifested.

At the instant t_1, the average distance between neighboring baryons equals 3×10^{-24} cm. They are thus separated by a distance much less than the Compton wavelength of a baryon, $\hbar/(mc) = 2 \times 10^{-14}$ cm. In the more exact formula for the equation of state, therefore,

$$\epsilon = n^2m^{-2} + n^{4/3} + mn,$$

the last two terms are less than the first by factors of 10^{20} and 10^{30}, respectively. This also shows the primary role played by the vecton field in the energy density.[7]

In a calculation of the entropy, the exact instant of time when the acoustic waves decay and transform into heat does not play a role. If the decay takes place later than this instant, when $t > t_1$, then an adiabatic decrease of the energy density of the acoustic waves takes place according to the law $\epsilon \sim n^{4/3}$, so $S = \epsilon^{3/4}/n = \text{const}$. Before the decay, the number of phonons in each mode of the acoustic oscillations is $\epsilon_{\text{acoustic}}/n\omega = bg^{-1/2} = 10^{15}$. After the decay, Planck equilibrium conditions exist. New degrees of freedom with a momentum of the order of the temperature T, i.e., with wavelengths of the order of $T^{-1} \ll \lambda_{\min} = n^{-1/3}$, are excited. These degrees of freedom correspond to transverse vecton modes, electromagnetic waves, e^+e^- pairs, $\nu\bar{\nu}$ pairs, and $B\bar{B}$ pairs. The energy of each degree of freedom then turns out on the average to be less than one quantum of energy of the field. It is absolutely less than one quantum for fermions and of the order of 1/20 of a quantum for bosons near the maximum of the Planck distribu-

7. It is just the small size of the baryon contribution compared to the energy of the vecton field that allows one not to take account of the density fluctuations resulting from the discreteness or "particulate" nature of the separate baryons.

tion and on the average. This is so even though $kT/(\hbar\omega) > 1$ in the Rayleigh-Jeans region.

The excitation of a large number of degrees of freedom with a simultaneous decrease in the number of elementary excitations per mode from 10^{15} to 1 constitutes the irreversible process whereby entropy arises from coherent oscillations. In the intermediate region between λ_{min} and large wavelengths, the perturbations whose amplitudes correspond to the parameter value $b \sim 10^{-4}$ decay without leaving observable traces. The hypothesis proposed specifies the magnitude of these perturbations, however.

Notice that the entropy arises at the instant t_1, but that the energy density and the pressure are still determined by the features of the vecton field. The thermal energy only constitutes a fraction b of the total energy. The transition to the RD situation takes place later, since $\epsilon_{vec} \sim n^2$ and $\epsilon_{ther} \sim n^{4/3}$. Hence $\epsilon_{vec} < \epsilon_{ther}$ beginning with the instant t_2, when $n < n_2 = b^{3/2}n_1 = 10^{-6}n_1$, so $t_2 > b^{-3/2}t_1 = 10^6 t_1 = 10^{-28}$ s. It remains unclear whether the metric perturbations are scalar (to use Lifshitz's terminology) or whether they satisfy the uniform-distribution principle (see § 12.7). In the latter case, the gravitational-wave density is comparable with (though less than) the density of electromagnetic radiation. The spectrum is then similar to a Planck spectrum and has a maximum for $\lambda \sim 100$ cm.

In conclusion, recall that the hypothesis set forth in this section so far makes use of unproved assumptions concerning the baryon equation of state and the perturbation spectrum. It is thus not required that we adopt this hypothesis.

One ought, therefore, to recall several points made by Grishchuk (1974; also § 24.17). According to him, the equations for gravitational waves are not conformally invariant, and as a consequence gravitons are also created at an isotropic singularity, i.e., near the instant $t = 0$ in a Friedmann solution. The spontaneous creation of gravitons is absent only in the special case when the scalar curvature R vanishes, and for this it is necessary that $P = \epsilon/3$.

This consideration imposes an important limitation on the equation of state in order to be consistent with the state of the Universe actually known from observations. In particular, the picture set forth earlier in this section, that of a cold, maximally stiff baryon gas, requires substantial revision.

Notice that the equation of state $P = \epsilon = n^2/m_p^2$ leads to abundant creation of gravitons, which ceases only after the entropy reaches the value $S = n_g/n_B \approx (Gm^2)^{-1/2} = 10^{18}$. This value vastly exceeds the observed value of $\sim 10^9$. If, however, the equation of state is not that of maximally stiff matter, then fewer numbers of gravitons are created; in particular there is no creation of gravitons at all if $P = \epsilon/3$ holds from the time of the singularity itself. If $P < \epsilon/3$, as for example in Hagedorn's model (see § 24.2), the role of the gravitons is of the order of unity near the singularity, but thereafter decreases in proportion to the amount of expansion and

becomes much less than unity when the hadron era is over. Thus, Grishchuk's considerations specifically excluded only the stiff equation of state.

Is this exclusion final? No. In his note, Grishchuk mentions the existence of modifications of GTR (which, however, coincide with standard GTR outside of a singularity) for which the creation of gravitons is forbidden in the isotropic case. Because the question was posed only very recently and is still not fully answered, we have presented here all possible considerations—including mutually exclusive ones.

24.9 The Steady-State Theory of the Universe

One of the methods for dealing with the problem of the singularity consists of avoiding the singularity altogether. Thus, remember that the proof that a singularity occurs is based on an assumption concerning the equation of state of the matter filling the Universe. By introducing a hypothetical field with a negative energy density, one can avoid a cosmological singularity.[8] A theory of this type, which is a theory of a stationary Universe, or, in the language of its authors, a steady-state theory, was developed by Hoyle (1948) and by Bondi and Gold (1948). Indeed, over the course of three decades, a significant fraction of the theoretical and observational investigations conducted have been devoted to verification, development, and criticism of the steady-state theory of the Universe. This reason alone is sufficient to oblige us to set forth the content of this theory and tentative opinions on it.

According to the steady-state theory, the metric of the Universe is

$$ds^2 = dt^2 - e^{2Ht}(dx^2 + dy^2 + dz^2) . \qquad (24.9.1)$$

It is apparent from the formula that the space (a t = const. section) is flat and that the Universe is homogeneous and isotropic. One can thus translate and rotate with respect to the spatial coordinates x, y, and z without changing observable features. These properties of the steady-state model are general properties of a Friedmann-Lemaître flat model.

The new property specific to the steady-state theory is the invariance of the metric under a group of translations with respect to time. All of the properties of the metric are the same at any moment t. On replacing t by $t + \tau$, we must merely redefine the coordinates via

$$x' = e^{H\tau}x , \quad y' = e^{H\tau}y \quad z' = e^{H\tau}z . \qquad (24.9.2)$$

The metric of the steady-state Universe also describes the expansion of the Universe and hence the redshift of spectral lines. The value of the Hubble parameter is H, the coefficient in the exponent of e^{2Ht}. At the same time, H turns out to be constant, not changing over the course of the

8. Conclusions concerning the collapse of separate, isolated bodies of large mass and high density are meanwhile preserved.

evolution.[9] Now the homogeneity of the Universe, i.e., the existence of a group of spatial translations that leaves the metric invariant, was codified by Bondi and Gold (1948) as the "cosmological principle." Generalizing this language, Bondi used the term "perfect cosmological principle" in referring to the invariance under a complete group of translations over space and time, a group that is characteristic of the metric of the steady-state Universe.

With the help of the GTR equations, let us find the stress-energy tensor, the right-hand side of the GTR equations, that is necessary in order that metric (24.9.1) hold with $H = \text{const}$. In correspondence with the general formulae of Part I, we obtain

$$\frac{1}{2}H^2 = \frac{4\pi G\epsilon}{3c^2}, \qquad H^2 = \frac{-4\pi G}{3c^2}(\epsilon + 3P), \tag{24.9.3}$$

from which it follows in particular that $\epsilon + 3P = -2\epsilon$ or $P = -\epsilon$. The quantities ϵ and P here include the contribution of the cosmological constant if it is nonzero: $\epsilon_\Lambda = [c^4/(8\pi G)]\Lambda$ and $P_\Lambda = -[c^4/(8\pi G)]/\Lambda$. Thus, in particular the metric of the steady-state Universe is an exact solution for an empty Universe with a positive cosmological constant[10] $\Lambda = 3H^2/c^2$.

But the authors of the steady-state theory propose their theory as a description of our actual, observable Universe, which contains matter in the form of stars, baryons, and RR. Since $P_m \ll \epsilon_m$ for ordinary matter, whose density is not less than 3×10^{-31} g cm^{-3}, then in the structure of the stress-energy tensor there must be a field, the "C-field," for which $P_c = -\epsilon_m - \epsilon_c$. Such a field violates one of the assumptions lying at the foundations of the theorem necessitating a singularity. A solution for which there is no singularity in the past or in the future therefore becomes formally possible.

A very important point is that the authors of the steady-state theory require that not only the metric and the geometry, but also the physics, be truly stationary. They do not confine themselves to a specification of the metric (24.9.1) in formulating the theory.

According to the steady-state theory of the Universe, then, it is necessary that the average density of matter, ρ_m, and of RR, ρ_r (the latter appended to the theory after 1965), be conserved over all time from $t = -\infty$. If the creation of baryons does not take place, then the equation $d\rho_m/dt = -3H\rho_m$, or $\rho_m \propto e^{-3Ht}$, holds for the density ρ_m. This is incompatible with the conceptual framework of the steady-state theory of the Universe, and so the authors make an additional assumption that baryons are created from the

9. One can write the Friedmann-Lemaitre metric in a form similar to that of metric (24.9.1) by replacing Ht by $\int H(t)\,dt$. The singularity of the Friedmann solution (in contrast to the steady-state one) is manifested in particular by the behavior $H \to \infty$ for some value of t, a value which one can choose to be zero.

10. See the formulae in § 4.2 concerning the limit $t \to \infty$.

C-field, thereby obtaining the equations

$$\frac{d\rho_m}{dt} = -3H\rho_m + A\,; \qquad \frac{d\epsilon_c}{dt} = -c^2 A - 3H(\epsilon_c + P_c)\,. \qquad (21.9.1)$$

By making a suitable choice of the value of A, namely $A = 3H\rho_m$, one can identically satisfy both of the conditions necessary for stationary character, $d\rho_m/dt = 0$ and $d\epsilon_c/dt = 0$, given the connection between[11] P_c and ϵ_c that follows from the GTR equations. However, it is then necessary to introduce a process unknown earlier whereby matter is created from the C-field. The steady-state theory of the Universe is for this reason also called a theory of continuous matter creation. The specific form of the A term and its dependence on ϵ_c and ρ_m or on other quantities remains indeterminate. One can suppose that $A = B\epsilon_c$, with B a constant, and consider ϵ_c to be independent of the coordinates. In this case, the creation occurs uniformly over all of space. Hoyle and Narlikar (1963), on the other hand, have studied the possibility that $A = B\epsilon_c\rho_m$, in which case the baryon creation takes place mainly in places where the matter density is already large. Independent of this point, though, the C-field is responsible for the baryon creation. Moreover, the term describing the creation must enter the equation for the matter with a positive sign and the equation for the energy of the C-field with a negative sign. Otherwise, one cannot make the solution compatible with the GTR equations.[12]

The perfect cosmological principle, a basic feature of the steady-state theory, is a quite stringent assertion. It states that the average density of galaxies, their average luminosity, the number and properties of radio sources, and so on, must be unchanging; these features must be the same in the distant past as they are at the present time. Thus, a definite log N–log S curve for the number of radio sources N versus flux S is predicted. It should not be surprising, therefore, that from the time of its inception, the steady-state theory of the Universe fell into the hands of astronomers, who subjected its cosmological inferences to observational verification. It turned out that the predicted dependence does not agree with the observations (see also § 8.3).

In evolutionary theories, one introduces an evolutionary function describing the time evolution of sources in order to make the theory agree with observations. But one cannot do this in the steady-state theory of the Universe because, though separate sources evolve, the Universe is, as a whole and on the average, stationary.

The steady-state theory of the Universe also predicts that the deceleration

11. The radiation density, ignored here, satisfies $\rho_r \ll \rho_m$. An account of it will not change the essence of the matter, however.

12. The underlying reason for this is that the GTR equation itself contains the energy-conservation law.

parameter q_0 is -1, another prediction which does not agree with observations.

It seems to us, however, that the primary and insurmountable difficulties associated with the steady-state theory are instead revealed in the attempt to specify at the level of elementary particles and field theory those hypotheses which must be introduced. The differential equation for $d\rho_m/dt$ can refer to the average density, but the creation of each separate baryon is a microscopic explosion.

If it is baryon-antibaryon pairs that are created, then annihilation must occur in a definite fraction of cases before any kind of mechanism acts to separate the baryons from the antibaryons spatially. Annihilation radiation has therefore been specifically sought, but it has not been found—even at a level 1000 times lower than that corresponding to what is expected from complete annihilation.

The creation of baryons without antibaryons would signify that the law of baryon-charge conservation is violated. According to general principles, consequently, the reverse process would also be possible, namely, the conversion of baryons into C-field energy. Such a process would be observed as the radioactive decay of nuclei that are stable in the ordinary sense.[13] It was mentioned, however, that such a process has not been observed in spite of special searches for it.

RR and its equilibrium Planck spectrum have a natural explanation in an evolutionary cosmology. The establishment of thermodynamic equilibrium takes place during the stage of the RD, dense plasma. In the steady-state theory, however, the Universe is transparent to radio waves even when the redshift z is very large, of the order of $z \sim 100$. The RR spectrum is therefore not amenable to an explanation in this theory.[14]

In view of the aggregate of observational data and theoretical considerations, the steady-state theory of the Universe can be considered refuted in 1976. One must, however, salute the intellectual boldness of its authors and keep in mind that the discussions concerning it were useful and promoted general progress in cosmology.

24.10 Mach's Principle and the Coincidences among Large Numbers Occurring in Physics and Cosmology

The relations between astronomy and physics constitute a complicated psychological combination. Thus, astronomers apply with relish whatever

13. See also Stanyukovich (1965) and discussion, Zel'dovich and Smorodinsky (1966*a*, *b*), and Stanyukovich (1966*a*).

14. Attempts to explain the 3 K radiation by the condensation of solid hydrogen have not survived criticism. In the pertinent work, Hoyle and Narlikar (1963) do not consider the optical properties of the hydrogen crystals, properties which differ significantly from those of blackbodies. To transform the condensation energy into radiation with a Planck spectrum, repeated scattering or absorption and reradiation are necessary.

is new in physics (experimentally or theoretically) to their science. These applications range from spectral analysis to photon counts and from classical mechanics to quantum physics. And while astronomers develop strong feelings about this borrowing of ideas—feelings not of anger, of course, but of appreciation—they think that nothing is as preposterous as the possibility of interfering actively in physics—not to take from physicists, but to give them something new! Yet astronomers have made splendid contributions along this path—the law of gravity and the determination of the speed of light, for example. Attempts to advance in the same direction and to give an astronomical explanation of the most important physical quantities and phenomena continue to this day. Indeed, those authors who originate these endeavors deserve great respect for their boldness in even posing the problems. Here we shall consider two connected ideas relevant to this theme—Mach's principle and the relation between the gravitational constant and the properties of the Universe as a whole.

Recall first that cosmology is that regime in which the laws of gravitation, quantum mechanics, and the theory of elementary particles simultaneously play a role. It was pointed out already that the constants G, $\hbar$, c, e, m_e or m_p characteristic of these theories could be used to make two similar dimensionless quantities

$$\frac{Gm^2}{\hbar c} = g \tag{24.10.1}$$

and

$$e^2/\hbar c = \alpha = (137.03)^{-1}\,. \tag{24.10.2}$$

The ratio of the electrostatic energy of the electron in the hydrogen atom to the rest energy of the electron is of the order of α^2. The fact that $\alpha < 1$ is just what allowed scientists to systematically create a nonrelativistic quantum theory of the atom, a good approximation indeed; and what thereafter allowed Dirac to develop the relativistic theory; and what then permitted others to develop quantum electrodynamics and renormalization theory. Were $\alpha > 1$ or $\alpha \sim 1$, the nonrelativistic theory, which is a first order approximation in α to the full relativistic theory, could not exist because all members of a series in α: α^n, α^{n+1}, α^{n+2}, . . . would be of comparable or successively increasing magnitude, with the result that the usual perturbation series in α would diverge and the physics would turn out to be infinitely difficult. Physicists have in fact encountered situations where a large dimensionless constant plays a role, i.e., strong-interaction physics and nuclear physics; and to this day these fields actually stand apart from atomic physics—perhaps not in a conceptual sense, but doubtless in the sense of the precision of their predictions.

Returning to astrophysics and gravitation, let us convince ourselves of the similarity between α and g. The Coulomb interaction between two particles involves an energy e^2/r, while the gravitational interaction in-

volves an energy Gm^2/r. Hence, Gm^2 plays the same role in gravity theory as e^2 plays in electrostatic theory, so they have the same dimensionality. We shall now substitute into the expression for g the mass of the proton and the mass of the electron:

$$g_p = \frac{Gm_p^2}{\hbar c} = 6 \times 10^{-39}\,; \qquad g_e = \frac{Gm_e^2}{\hbar c} = 2 \times 10^{-45}\,. \tag{24.10.3}$$

In both cases we obtain an extraordinarily small quantity. A natural measure of such small quantities is given by the natural logarithm:

$$-\ln g_p = 87.9\,; \qquad -\ln g_e = 102.8\,. \tag{24.10.4}$$

Hence, even though the logarithm of g differs strongly from unity, the logarithms for the two particles, the proton and the electron, do not differ sharply from each other. In the subsequent development, we shall not distinguish between these values; we shall therefore take the rounded-off values

$$-\ln g = 92\,, \qquad g = 10^{-40}\,. \tag{24.10.5}$$

Among physicists the conviction exists that dimensionless quantities differing substantially from unity must be explained and that these numbers are subject matter of a qualitative theory. This conviction further suggests that the closeness in value of large dimensionless numbers[15] arising in theories of various natural phenomena indicates the presence of intrinsic connections among these phenomena. Such connections, it is believed, can then serve as a beacon illuminating the path along which science should develop.

Landau has given an example of such an approach. He has pointed out that $\alpha \sim (\ln g)^{-1}$. But there exists in addition a conceptual basis underlying this coincidence. In a definite variant of quantum electrodynamics, the observable charge of an elementary particle is limited by an inequality in which a cutoff momentum enters as the argument of a logarithm in the denominator of the limiting expression. This cutoff represents the maximum momentum for which the theory is valid. In this sense, electrodynamics without a limiting momentum is incomplete and not logically closed, for it leads to the conclusion that the observable charge necessary equals zero. In this context Landau interprets the relation between α and g as showing that it is gravitational effects that limit the maximum momenta to be considered. Thus, electrodynamics exists in a strict sense only given a simultaneous account of the role of gravitation.[16]

15. For brevity, we shall speak everywhere of "large" numbers. If necessary, this means we must invert small numbers: $\hbar c/e^2 = \alpha^{-1} = 137$; $\hbar c/(Gm^2) = g^{-1} = 10^{40}$.

16. It should be pointed out that in the last few years another approach has been discussed, intimately tying electrodynamics with weak interaction theory (Weinberg, Salam, and many others). This approach does not need to consider gravitation in order to be logically sound.

As another example of the use of large numbers relating to astronomy, we note the estimate of the maximum mass of a white dwarf star (the Chandrasekhar limit). Salpeter has shown that the maximum number of nucleons in a white dwarf N equals $g_p^{-3/2}$. The quantity g_p enters into this calculation, which applies to nucleons (protons), because the electrons are relativistic and their mass must not enter into the result. Numerical factors such as 2, 3, and π are omitted, as are ratios of the atomic weights of nuclei to their atomic masses. Thanks to favorable cancellation, the numerical agreement is not bad—$M = 1.2\,M_\odot$ corresponds to 1.5×10^{57} baryons as compared to $g_p^{-3/2} = 2 \times 10^{57}$. The same dimensional argument could be applied to the neutron star mass limit, and we know already that this limit is of the order of 2 or 3 $M_\odot$. Here we use also the fact that the rest masses of all strongly interacting particles are of the same order of magnitude.

Still earlier, however, Dirac boldly applied the "large numbers" idea to cosmology itself. Indeed, this application will be examined in detail below. Note though that in the absence of a logically closed theory, hints or indications concerning the role of large numbers are of course not unique.

One direction of research consists of inferring the properties of the Universe from the theory of gravitation, i.e., under the assumption that the quantity g (measurable locally and being part of laboratory physics) is basic to cosmology. There is, however, another approach—one based on the hypothesis that the Universe as a whole has a definite role in determining the local laws of physics. This hypothesis is called Mach's principle.

In its embryonic form, Mach's principle consists of a question relevant to Newtonian physics: Why is there centrifugal force in a rotating body? The following related question then arises: Is rotation absolute, or is it rotation relative to the entire remaining Universe that is important? Experiment shows that the rotation of a body relative to the motionless, distant stars indeed leads to the appearance of centrifugal force. The question then becomes whether there are centrifugal forces in a world in which there are neither distant or nearby stars nor, in fact, any matter except for the body under consideration and, perhaps, one observer.

On the basis of direct experiment, one can evidently not answer this last question because there is no such Universe. Any such experiment must forever be an imaginary or thought experiment. Mach's assumption is that no "absolute" rotation exists: no inertial (particularly no centrifugal) forces would exist in the empty Universe. But Mach himself never worked this assumption into quantitative theory.

In GTR, the forces of inertia and of gravity are closely connected with each other. It is therefore not surprising that with Mach's principle as a motivation, the question subsequently arose of whether or not gravity depends on properties of the Universe as a whole. If it does, then the quantity g that characterizes gravity can depend on the large numbers characterizing

the properties of the Universe as a whole. Mach's principle might therefore have something in common with the problem of the role of large numbers in local laboratory physics. It is by such reasoning that a group of questions is obtained which falls under the general heading of the "many faces of Mach."

We shall now consider in succession two formulations of the question of the role of large numbers. We shall begin with the factual side of the matter.

First, the one large number $g^{-1} = 10^{40}$ was found above to characterize local physics. To go along with this, Dirac considers the total number of nucleons in the Universe N as a number characterizing the Universe. He obtains this number by taking the radius of the Universe to be $c/H = 10^{28}$ cm. Since the critical density of 10^{-28} g cm^{-3} corresponds to a particle number density $n = 10^{-5}$ cm^{-3}, we have that

$$N = (4\pi/3)(c/H)^3 n = 4 \times 10^{79} \approx g^{-2}. \tag{24.10.6}$$

This is a remarkable coincidence, which Dirac considers too good to be accidental.

Let us secondly substitute the expression for the critical density, $\rho_c = 3H^2/(8\pi G)$, and the corresponding expression for the number density of nucleons, $n_c = 3H^2/(8\pi G m_p)$, from which $N \sim c^3/(HGm_p)$ follows, into the relation $N = g^{-2}$. By omitting numerical factors and by performing some simple algebra, we obtain another remarkable relation that is equivalent to the preceding relation.

$$H = \frac{m_p c^2}{\hbar} \frac{G m_p^2}{\hbar c} = g \frac{m_p c^2}{\hbar}, \tag{24.10.7a}$$

$$t = H^{-1} = \frac{\hbar}{m_p c^2} g^{-1}. \tag{24.10.7b}$$

See Gamow (1948, 1968) and Stanyukovich (1965). The coincidence changes little if, instead of g, one takes the ratio $Gm_p^2/e^2 = 0.8 \times 10^{-36}$, as is done by Dirac in the original work. For modern work see Dirac (1974). Gamow calls the unit of time characterizing the nucleon, $\hbar/(m_p c^2) = 10^{-24}$ s, a "tempon." The age of the Universe as measured in tempons is then $\sim 3 \times 10^{41}$. This is a dimensionless number, indeed, one that reminds us of the quantity $g^{-1} = 10^{40}$ characteristic of the local effects of gravitation.

We shall now make the physical ideas connected with these coincidences more precise. First of all, the very idea of a total number of nucleons makes sense only for a closed Universe. To be more precise, one should take $\rho = \rho_c \Omega$ with $\Omega > 1$, the curvature radius $a = cH^{-1} f(\Omega)$ and the volume $V = 2\pi^2 a^3$ (see Part I).

The assumption of a connection between the total baryon number N in the closed Universe and the constant g is possible; it does not lead to variability of gravitational constants or masses of the elementary particles. Of course, in this case there is a hidden assumption that even the baryons

which are outside the particle horizon are still influencing the local physics. The connection between H or t and g is now only a temporary coincidence approximately valid for the part of the closed Universe evolution near maximum expansion. But in the case of an open Universe the overall number of baryons is infinite. The radius of curvature of the hyperbolic universe goes to infinity as $1 - \Omega$ tends to zero. The best interpretation of (24.10.6) for the case of an open Universe is the idea that N is the number of baryons inside the particle horizon which influence the local physics in a Universe of infinite overall size.

But in this case we can no longer leave all physical quantities constant. In the simplest case of a flat, matter-filled Universe with constant G, one obtains $H \sim t^{-1}$, $n \propto \rho \sim Gt^{-2}$, so the right-hand side of (24.10.6) is proportional to t. In order to preserve (24.10.6) one needs to assume that G (perhaps and/or other constants also) varies with t. But of course all the beautiful edifice of theoretical physics is ruined. And there are no other arguments in support of such a conception except the aforementioned coincidence between N and g^{-2}.

The possible variability of G and the consequences thereof have been widely discussed. Thus, in the variant of the theory in which only G varies, it follows from formula (24.10.7) that G was 1.5 to 2 times larger 5×10^9 yr ago than it is today. This time period is roughly the age of the solar system and therefore of the Earth. Given then the conservation of angular momentum for the motion of the Earth around the Sun, a larger value of G corresponds to a distance of the Earth from the Sun that is smaller by a factor of 1.5 to 2. By the theory of the internal structure of stars, the luminosity of the Sun also depends on G—indeed, quite strongly, roughly as G^7. Thus, the flux of solar heat falling on a unit of area of the Earth varies as G^9 in any theory in which G is variable. Geological and paleontological evidence decisively contradict this.

Finally, there exists an attempt at reading the formula for G in such a way that the specific traits of the contemporary epoch—life, mankind, civilization, and the very possibility of studying nature and writing formulae and books (including this book!)—are allowed only when H falls to a value of the order of $g(m_p c^2)/\hbar$. Thus, the formation of galaxies and stars can take place in the hot Universe only after the recombination of protons and electrons, i.e., after the end of the RD stage. Now this time (given $\Omega = 1$) corresponds to the cosmological time $t_r = 5 \times 10^{12}$ s or to an age of 5×10^{36} as expressed in units of tempons, a time that also does not differ significantly from g^{-1} or g_p^{-1} or from $e^2/(Gm_p^2) = 1.25 \times 10^{36}$. It was Gamow who pointed out this coincidence.

Such an assertion is wholly compatible with ordinary physics—with the laws of nature—in which the constants G, $\hbar$, and c and the masses of stars and of elementary particles are constant. To see this, note that relation (24.10.7a) in reality is an assertion applying to the specific entropy of the

matter filling the hot Universe. But the specific entropy is a local quantity; one can define it both for an open and for a closed Universe (in distinction to the situation for the total number of baryons N). Thus, it is worthwhile to distinguish relation (24.10.7a) from relation (24.10.6) if it is read as a condition for the appearance of a definite era in the evolution.

Numerically, relation (24.10.7a) is equivalent (roughly, of course) to the relation $S = n_\gamma/n_B = [Gm^2/(\hbar c)]^{-1/4} = g^{-1/4}$. The latter was obtained by Zel'dovich and Novikov (1967*b*) as a deduction from a scheme whose premises are at the present time improbable. As an empirical formula, however, the assertion remains true.

In a theory with a nonzero cosmological constant Λ (see chap. 4), one can assume, without violating the "principle of the constancy of the constants," that $\epsilon_\Lambda \approx m^6Gc^4/\hbar^4$. In the model investigated by Lemaître, Petrosian, Salpeter, Shklovsky, and Kardashev (see § 4.2), $\epsilon_\Lambda < 0$, and a period of delayed expansion is obtained. The age of the Universe at this instant and the characteristic time of the subsequent expansion satisfy the relation $\rho \sim \epsilon_\Lambda/c^2 \sim 1/Gt^2$ leading to Gamow's (24.10.7b) (Zel'dovich 1968). In recent years, however, those first observational data (the concentration of quasars at a redshift $z = 1.95$) on which these authors based the revival of a model with a delayed expansion and therefore with $\Lambda \neq 0$ have not been confirmed.

Let us now return to our general analysis of Mach's many-faced principle. The essence of the principle is the assertion that the inertia of a body is determined by its (gravitational-inertial) interaction with other bodies in the Universe. This principle played a large role in the creation of GTR by Einstein. But after the creation of the theory of relativity, it became clear that Mach's principle is not contained in it! Let us consider this question in somewhat more detail.

What is the meaning of the assertion that inertia is determined by an interaction with other bodies? A straightforward answer to the question is this: The inertial mass of a body, i.e., the measure of its resistance to an applied force, is determined by its interaction with other bodies. If these bodies did not exist, then a test body would have no inertial mass. On the other hand, "the inertia of a body must increase when ponderable masses are piled up in its neighborhood" (Einstein 1966).

Neither of the last assertions is fulfilled in the theory of relativity. In the first place, the special theory of relativity is valid in empty space, in which a body possesses inertia and in which there are Coriolis and centrifugal forces in a rotating system. In the second place, a given force—that of a compressed spring, for example—always transmits the same acceleration to a given body independent of the proximity of heavy masses or of the absence of them. Hence, the inertia of a body does not grow "when ponderable masses are piled up in its neighborhood." The opposite assertion by Einstein is connected with an erroneous interpretation of a formula he

obtained. (Brans 1962 and Dicke 1962 have pointed out this error.) In this sense, confirmation of the theory of relativity is a blow to Mach's principle.

The situation is different regarding several other physical ideas which are sometimes connected with Mach's principle. Thus, Einstein (1966) writes that, from the point of view of Mach, one ought to expect that "a body must experience an accelerating force when neighboring masses are accelerated, and, in fact, this force must be in the same direction as that acceleration." And furthermore, "a rotating hollow body must generate inside of itself 'a Coriolis field,' which deflects moving bodies in the sense of the rotation and a radial centrifugal field as well."

Both effects exist in GTR (see Vol. 1). But they are not connected with a change in the inertial properties of a test body. They instead describe a change in an inertial reference system due to the motion of gravitating masses. Said otherwise, the mass of the test body remains unchanged, but the inertial system of reference is dependent on the presence and motion of neighboring bodies. These effects, which were pointed out by Einstein, are theoretically of the same character as is the change in an inertial system of reference in the presence of a motionless gravitating mass (an inertial system near a nonrotating planet, let us say). The inertial system falls freely in the gravitational field of the mass, whereas it coincides with an inertial system at infinity in the absence of the mass. It is evident that such a change in the inertial system is in no way connected with a change in the inertial properties of a body. Recall again that such inertial properties constitute a measure of the resistance of the body to accelerating (nongravitational) forces.

Having thus reclassified these two other effects as non-Machian, we conclude that Mach's principle does not exist in GTR.[17]

Another viewpoint regarding Mach's principle is more radical. In holding to this view, its proponents consider GTR to be incorrect or incomplete precisely because the action of distant masses is not therein accounted for in a clear form (they desire something more than an influence on the metric). These investigators consider the existence of a solution of the field equations for empty space which gives observed inertial properties (compare with the remarks made above) as a shortcoming of GTR which a future theory can overcome.

Of course, a debate over a future theory is always difficult. It tests the sense of humor and the integrity of the arguing parties. It is dangerous to find oneself in the role of the critic Latonsky, who wrote the article "Let's Strike a Blow against Pilatovshchin," wherein the Master's[18] unpublished novel *Pilate and Jesus Christ* is attacked.

17. Concerning Mach's principle see also Wheeler (1964); Misner, Thorne, and Wheeler (1973).

18. See M. Bulgakov's *The Master and Margarita*. Bulgakov (1888–1970), an outstanding Russian writer, depicted the Civil War and Soviet society during the period between World Wars I and II.

And is it not all the same regarding both an attack on unborn theories and a defense of GTR? One is thus also obliged to note the facts that are unfavorable for the critics of GTR. In short, GTR has encountered neither experimental refutation nor logical difficulties. There are therefore no realistic, objective reasons impelling one to replace it by a "more Machian theory."

If, however, one turns to the Universe in which we live, then the distant stars and the RR in it actually distinguish at each point one system of reference at rest. Not only rotational motion, but also translational motion relative to the distant stars, can be observed and measured.

Thus, in the rest system we observe in all directions a redshift of the light from the distant galaxies. We also observe that the temperature of the radio background in all directions equals 2.7 K. In a system undergoing uniform translational motion though, we would observe a relative blueshift in the direction of motion. The background temperature would meanwhile be higher than 2.7 K in that direction and lower in the opposite direction. Mach's principle would necessarily return us not only to Newton—to teach us how to distinguish rotation in the Universe—but also to Aristotle—to teach us how to determine absolute rest in the Universe.

If experiment shows that the laws of nature are Lorentz invariant, then this directly demonstrates—independent of what the laws are—the absence of influence of distant bodies. There is every reason to believe that the laws are indeed Lorentz invariant, and in this sense the attractiveness of the explanation of a particular group of phenomena by the action of distant bodies is strongly undermined.

In summary, we do not regard as probable the radical revision of GTR in the direction of Mach's ideas. We moreover consider the problem of the large numbers to be realistic only in a formulation which connects local properties and which therefore does not necessitate the variability of physical quantities—in particular, variability of the fundamental physical constants.

24.11 The General Theory of Relativity (GTR) and the Topological Structure of the Universe

The question of the topological structure of the Universe is one of the most important problems lying at the border between the natural sciences and philosophy. This question received a new meaning and new content when, at the beginning of the century, the theory of relativity was created.

Before this, it was considered self-evident that space is three-dimensional and Euclidean and that time is an independent variable. It was the special theory of relativity (STR) that changed this naive (from today's point of view) belief. According to STR, space and time turn out to be combined into a single four-dimensional manifold. Indeed, a transformation law

giving the transformation of coordinates and of time upon a transition between relatively moving systems of coordinates was established as the concrete expression of this union. These transformations were first found by Poincaré from the condition that Maxwell's equation for the electromagnetic field be invariant. It is customary, however, to speak of them as Lorentz transformations; one thereby takes note of Lorentz's role in the development of the theory of the electron and of electromagnetism. Einstein's immortal contribution, by virtue of which he, in particular, is justly recognized as the creator of the theory of relativity, consists of the realization that the connection between space and time is universal. Einstein understood that the question is not of the properties of a single special field (the electromagnetic field), but rather of the general properties of all of nature, of all particles and fields, because it concerns space and time common to all the fields. It was not by chance that only Einstein expressed the idea of the equivalence of mass and energy through the formula $E = mc^2$ and that, without wavering, he applied this relation to the atomic nucleus in spite of the absence of a concrete theory of nuclear forces.

But let us return now to the question of the structure of the world. The creation of STR was of surprisingly little value with regard to this question. With its development there arose the idea of a four-dimensional, pseudo-Euclidean manifold (Minkowski space), in which one can choose both the time axis and the corresponding, orthogonal, three-dimensional manifold—"space"—in many ways. In this picture, however, there first of all arises no doubt that the time t and each of the three spatial Cartesian coordinates might change their character in the interval from $-\infty$ to ∞. The special properties of Minkowski space (the existence of the Lorentz transformation) lead to the preservation of the concepts of an absolute past and an absolute future. No special question concerning the topology of spacetime arises. Before the creation of GTR, therefore, the problem of the structure of the Universe was understood as simply the question of the spatial distribution and evolution of the matter and fields filling the Universe. The structure of the space itself seemed self-evident.

In the schematic account above, we have remained silent about the views of Lobachevsky, Bol'yai, and Riemann and about the unpublished ideas of Gauss. The deep ideas of these creators of non-Euclidean geometry anticipated GTR; we shall discuss their ideas in the context of GTR. Now it is just GTR which poses the question of the structure of spacetime itself. *Locally*, the question is formulated as a question concerning the metric and the geometry of four-dimensional spacetime. The three-dimensionality of space and the four-dimensionality and continuity (differentiability) of spacetime are accepted as a unifying deduction based upon experiment. From this alone STR flows with logical inevitability as the correct theory applicable on small scales.

The history of science shows that any transition to a new area of research

leads to the formulation of new laws and, what is more important, new concepts. Textbook examples of such new areas include high-speed phenomena, which led to STR, and phenomena on small, atomic scales, which led to quantum mechanics. It was just the transition to the study of phenomena taking place on large scales that led to the creation of GTR.

This is not the place to dwell on a peculiar intermediate development—the theory of gravity in a narrow sense. For the matter density realizable in nature, gravitational effects can successfully be observed over even small spatial scales thanks to the high sensitivity of experiment. So, for example, in Cavendish's experiments over laboratory scales, interaction effects are observed for which the associated energy is a fraction of the order of 10^{-25} of the rest energy mc^2 involved. In this intermediate region, one can make use of the linearized theory of gravitation; complete GTR is not necessary.

However, let us turn to the scale over which astronomical research is carried out, a scale of the order of 10^{28} cm. On this scale, that of the Universe itself, the energy of the gravitational interaction is of the order of the rest mass involved even though the average density of matter in the Universe is 30 orders of magnitude smaller than the density of the lead spheres Cavendish employed in his experiment. Correspondingly, the curvature of space predicted by GTR is of the order of unity.[19]

It is just the possibility of a large spacetime curvature, i.e., the substantial deviation of spacetime's *global* properties from Euclidean properties, that naturally and inevitably leads to the question of the *topology* of spacetime.

For the particular case of a homogeneous, isotropic model, we have already encountered this question. Thus, given a density greater than the critical density, $\Omega > 1$, the metric takes the form

$$ds^2 = dt^2 - a^2(t)[dr^2 + \sin^2 r(d\theta^2 + \sin^2 \theta d\phi^2)] .$$

The invariant dimensionless measure of the curvature is the product of the curvature (with the dimensions cm^{-2}) and the square of the characteristic scale (with dimensions cm^2). One can formulate the situation in this fashion. Local studies involve observations in that system of reference in which the observer is at rest relative to the surrounding matter and radiation. These observations show that the Universe is homogeneous (that the properties of the Universe are independent of translation in the three-dimensional space) and isotropic (that the properties are independent of rotations in this space).[20] Furthermore, measurements of the matter density, of gravitational effects, and, in principle, of the metric in our neighborhood give the curvature of the chosen three-dimensional space. Assume here that the measurements give $\Omega > 1$ with certainty, even though the accuracy of the

19. So are the effects in neutron stars and black holes.

20. A transformation to a moving system (a Lorentz transformation) would destroy the homogeneity and isotropy of the matter distribution and of the radiation fluxes.

relevant measurements is today insufficient to warrant drawing such a conclusion and even though $\Omega < 1$ seems more probable.

A conclusion that the Universe is closed is then obtained on the basis of an extrapolation. It is simply necessary to extrapolate the properties of the metric in our neighborhood to distant regions by making use of the locally established homogeneity of the Universe.

For the metric written above, local observations plus homogeneity lead to the conclusion that the coefficient for the angular part, $d\theta^2 + \sin^2\theta d\phi^2$, equals $a^2(t)\sin^2 r$. But the function $\sin r$ goes to zero for $r = \pi$, so the metric is not extensible to values $r > \pi$. It correspondingly follows that the total volume is $V = 2\pi^2 a^3$ and that all other consequences are those characteristic of a closed Universe. Such is the logical chain which necessarily leads to the conclusion that a Universe with $\Omega > 1$ is closed.

Hence, there is indeed a case when a non-Euclidean topology is necessary. A further natural generalization follows; one can look for models with a non-Euclidean topology even when such a topology is not necessary.

We now shall present a very simple example of this data. Suppose that the local studies yield $\Omega = 1$ and, correspondingly, the flat three-dimensional metric

$$\begin{aligned} ds^2 &= dt^2 - b^2(t)[dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2)] \\ &= dt^2 - b^2(t)[dx^2 + dy^2 + dz^2] . \end{aligned}$$

This metric is evidently compatible with a Euclidean topology, since $-\infty \leq x \leq \infty$, $-\infty \leq y \leq \infty$, and $-\infty \leq z \leq \infty$ is a possible set of coordinate values. But the topology of a closed three-dimensional space is also possible, although not necessary. In this case, we make the identifications

$$\begin{aligned} (x + k, y, z) &\equiv (x, y, z) , \\ (x, y + p, z) &\equiv (x, y, z) , \\ (x, y, z + n) &\equiv (x, y, z) . \end{aligned}$$

In the comoving three-dimensional space, a parallelepiped type of periodicity with a volume per period of kpn [or $b^3(t)kpn$ in the instantaneous physical three-space] is distinguished. The space is then said to have the structure of a "three-dimensional torus."

During the nineteenth century, we would have spoken of a Universe with periodic structure. We would now say instead that a flat, closed, three-dimensional space exists. From the observational point of view, though, the two formulations are similar. For, given a spatial period sufficiently small that $b(t)k$, $b(t)p$, and $b(t)n$ are smaller than the path length ct traversed by light during the evolution of the Universe to the time t, then one can see the same object (a galaxy, for example) by looking in diverse directions. It is necessary to emphasize that from the old (nineteenth century) point

of view, we would be talking about identical objects with identical properties. From the contemporary, topological point of view, it is assumed that all of space has a limited volume—that it is closed—and that the same exact object is being seen more than once. This involves a change in the global law of light propagation (and the laws describing any other motions) to correspond with this condition of image identification. Note, though, that the local law of propagation is nowhere changed and that the world model so constructed has no boundaries or gaps. Observations would then give many simultaneously visible "ghost" images corresponding to each object of a given type. However, these simultaneous images visible in different directions are, on account of the properties of light propagation, emitted by the object at various instants of time. In principle, the situation does not differ from the predictions of the theory of a spherical closed model, e.g., the Lemaître model with a cosmological constant (see § 4.1 and § 4.2). Remember, however, that special searches for such matching objects in opposite sections of the sky, searches undertaken in connection with this model, have not led to a positive result.

Though the aforementioned necessity to consider topology when studying cosmological models is based on the overall history and philosophy of science, the results of a topological analysis are nontrivial. This is what the simple examples examined above illustrate. Now the number of different topological variants of a four-dimensional spacetime is extraordinarily large, so at first glance an enormous number of possibilities is open in cosmology. In reality, though, even minimal information concerning the actual properties of the Universe sharply limits the immense scope of topological structure allowed. But, and this is particularly important, the theory describing the evolution of the Universe turns out to be almost independent of the topology. It is just this fact which explains the comparatively small number of works devoted to the question. Note, however, the studies by Süveges (1966), Paal (1971), Zel'dovich and Novikov (1967*a*, *b*), Sokolov (1970), Sokolov and Shvartsman (1974), and Sokolov and Starobinsky (1975).

We shall now examine some of the constraints imposed by observations on the possible topology of a model Universe. We shall be considering first a spatially homogeneous model and thereafter a weakly perturbed, spatially homogeneous model. Notice first of all that the conditions of spatial homogeneity distinguishes parallel, three-dimensional manifolds (spaces at different instants of time) on which the homogeneity is realized and the one definite time axis (see above) orthogonal to these manifolds. Along the time axis, identifications as described above are forbidden by the laws of physics and are therefore excluded. This ban is ordinarily based upon the principle of causality. Thus, upon the identification of t_2 with t_1, where $t_2 > t_1$, it turns out that the future and the past are entangled. Interference in a physical phenomenon at the moment t_2 and the point x gives rise, by

this identification, to changes at the moment t_1 and the point x. As time evolves from t_1, consequently, the original interference affects events at some time $t_3 > t_1$. If, in addition, $t_2 > t_3$, the net result is that the past has been influenced by the future. According to the principle of causality, such an influence cannot occur.

The application of the causality principle is difficult if one carries out an identification in time with a simultaneous translation along one or several spatial coordinates, i.e., if one identifies the world points (t_2, x) and $(t_1, x + k)$. However, with due regard for spatial homogeneity and with an account of the irreversibility of evolution, one can also forbid any such world-point identifications that include a coordinate translation. This is because the specific entropy of the Universe (per single baryon or per unit volume of the comoving space) grows monotonically over the course of time. Also, the average abundance of protons in the matter of the Universe monotonically decreases as a result of nuclear reactions. Thus, for these reasons, the situation at an instant of time t_2 is not identical with the situation at the earlier instant of time t_1.

With respect to the time axis, consequently, there remains only the question of whether the scale extends (1) from $-\infty$ to $+\infty$, (2) from some initial time t_1 to $+\infty$, (3) from $-\infty$ to some final time t_2, or (4) from t_1 to t_2. For homogeneous models, the answer to this question is obtained by means of an integration of the equations of evolution. In principle, there is no arbitrariness here. At the present time, however, observational data, while refuting possibilities (1) and (3), do not permit one to choose between possibilities (2) and (4). A singular state is, of course, what is meant here by an initial time or a final time for the time scale.

Hence, the question of the topology of the Universe narrows to the question of the topology of the three-dimensional space, in which case there still remain a significant number of variants. With reference to geometry (see Efimov 1961; Wolf 1967; Süveges 1966), 18 topological variants of flat, three-dimensional space have been counted, as well as an infinite (!) number of topological variants of hyperboloid, isotropic three-dimensional spaces with constant negative curvature.

Any anisotropy in a Universe that is still homogeneous additionally limits the number of possible topological variants. To see this, consider an identification of two surfaces in the three-dimensional space. Make a circumnavigation of each spatial cycle characterizing the topological variant. The coincidence of the values of scalar quantities before and after the circumnavigation is guaranteed by the condition of homogeneity.[21] But

21. As is well known, the laws of physics are not invariant with respect to mirror reflection (parity inversion) in three-dimensional space. It is therefore necessary to distinguish between scalars and pseudoscalars. Equivalently, a circumnavigation over a cycle must not transform a right-handed system into a left-handed system. This question was considered by Süveges (1966) and by Zel'dovich and Novikov (1967*a*, *b*).

after such a circumnavigation in an anisotropic Universe, it is also necessary to return to the original point with the same values for vector and tensor quantities and with the same directions for the velocities and the principal deformation axes. These requirements are more stringent than are those for an isotropic Universe. The number of possible topological variants is therefore smaller.

The remaining number of possible topological variants depends on the nature of three-dimensional space (in the sense of the Bianchi classification). In a flat space, for example, the possibility of a transition to the topology of a cylinder or of a torus continues to hold. This possibility corresponds to a periodic identification condition. Note finally that of all of the cosmological solutions of Einstein's equations, the Friedmann solution admits the maximum number of topological variants.

We shall now turn to the evolutionary aspect of the topology problem. The most important facts relevant here are that in a homogeneous cosmology (1) the topology does not change during the course of the evolution, and (2) the equations for local quantities do not depend on the particular topological variant employed. Remember that one can express all local properties in terms of the values of only three groups of quantities. These are such physical quantities as the density and the other entries of the stress-energy tensor for matter and fields, such geometrical quantities as the curvature tensor, and the structure constants characterizing the spatial anisotropy. The last are somewhat similar to the curvature in a space of Killing vectors. Consequently, the analysis of how nuclear reactions and many other analogous processes proceed in a homogeneous Universe—or, more precisely, in all anisotropic variants of a homogeneous model of the Universe—turns out to be strictly independent of the topological properties of the model. The topology of the models does, however, subsequently affect the solution of a more general, inhomogeneous problem. One can study various aspects of the problem.

One of them is connected with the mixmaster model of the Universe. People have studied ranges of influence and the possibility of information exchange between diverse parts of the Universe. It is evident that identifications of the type that connect the points x and $x + k$ in a flat Universe lead to the possibility of complete information exchange. A topological variant that includes the identification of otherwise distant points decreases the size of the Universe; thereby, mixing becomes possible even in a locally Friedmann Universe. In this case, though, the absolute value of the identification length k and the specific character of those inhomogeneities which are faced with eventual smoothing out are important in any discussion of the evolution of the model. A detailed discussion of this now would be premature, however.

Another approach to the effect of topology on the evolution of an inhomogeneous Universe is possible on the lines of perturbation theory accord-

ing to which small perturbations can be decomposed into characteristic functions. The choice of the characteristic functions and their spectrum depends radically on the topology of the space. To see this clearly, turn again to the elementary example of flat space. Given a Euclidean topology, the spectrum is continuous; the characteristic functions are essentially $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$, with any value of the wave vector $\boldsymbol{k}$ satisfying $|\boldsymbol{k}| \to 0$, or any value of the wavelength λ satisfying $\lambda \to \infty$, admissible. Now in any space with a point identification of the torus type, the perturbations must also obey the identification condition. This means that the wave vector must have components that are multiples of the inverse of the periodicity lengths. The perturbations are therefore decomposed into a Fourier series with a discrete spectrum instead of a Fourier integral with a continuous spectrum (the latter is valid in the case of Euclidean space only). Such a result is entirely natural if one keeps in mind that the observable properties of a space of the torus type are equivalent to those of a Euclidean space with a periodic distribution of all quantities.

Now it is well known that the amplitude of large-wavelength perturbations is small in the real Universe. This fact has been established observationally on the basis of the small size of RR fluctuations. One of the problems facing a future theory is to obtain this property of the perturbations from first principles. Perhaps one can connect this with topology, with the property that there are no long-wavelength fluctuations in models with identifications of nearby points.

At the present time, the scale of, for example, galaxy clusters is not greater than 40–100 Mpc. On larger scales, the inhomogeneity amplitude falls quite noticeably. In recent work, Sokolov and Shvartsman (1974) maintain that these observational data do not exclude a topological structure wherein the elementary periodic cell has a linear scale (at the present time) of the order of 400 Mpc in length. On the scale of distances corresponding to a Hubble parameter $H = 75\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$, such a periodicity scale corresponds to a redshift of $z = 0.1$. If one seriously admits such a contraction in the scales of the Universe, then the large-scale RR isotropy has a new, unexpected explanation.

Though interesting, one must actually consider Sokolov and Shvartsman's assertion as an appeal to astronomers to carry out observations that are specifically oriented toward a proof that the Universe is not small. This has not been proved at the present time, but it can be proved if data are accumulated concerning radio sources (= quasars) that are visible out to a redshift of $z \sim 3$–4, according to estimates. Unfortunately, a problem here is that the active lifetime of radio sources can be small. This can make the analysis of the observations difficult by decreasing the number of objects whose images one can simultaneously record by means of rays incident along diverse paths.

It is impossible and unscientific to determine the "probability" that the

Universe is topologically complicated. But it would be quite desirable to organize carefully done observations in such a way as to help explain this point. Think of the enormous significance there would be in a discovery that the Universe is topologically finite!

We merely mention the connection between topology and quantum phenomena connected with the change of the spectrum of eigenfunctions of all fields. The reader will find these topics in Starobinsky (1975) and Ford (1975).

24.12 Local Topology, White Holes, Black Holes, and Cosmology

The possibility that strange topological structures exist in GTR has been used by several authors for an explanation of astronomical phenomena. Already in 1928, Jeans wrote in connection with the question of the nature of the spiral arms of galaxies: "The type of conjecture which presents itself, somewhat insistently, is that the centers of the nebulae are of the nature of 'singular points,' at which matter is poured into our Universe from some other, and entirely extraneous, spatial dimension, so that, to a denizen of our Universe, they appear as points at which matter is being continually created."

A still greater temptation is to use the inflow of material and energy from the "other dimension" to explain the activity of galactic nuclei and quasars. Can one also make use of this inflow in the steady-state theory of the Universe? To answer these questions, we shall now explain what constraints GTR imposes.

Let us surround a three-dimensional region in which a local singularity is assumed by a distant, two-dimensional closed surface—a spherical surface, for example. The surface itself lies in a region where the gravitational field is weak. By studying the spatial metric outside this surface, one can determine the total mass found inside (and also the angular momentum) and, by measuring the electric field, the total charge too. Now the GTR equations written for the external, nonsingular region lead to conservation laws which express the fact that the mass located inside the enclosed region changes only when matter or energy, including that in the form of gravitational waves, flows into or out of the region through the surface. Hence, independent of the most fantastic possible assumptions concerning the nature of the singularity inside the surface, one can say in advance how much mass and energy one can obtain from the enclosed region.[22]

Thus, Jeans's graphic picture—the opening to another Universe from which energy can enter in any amount—turns out to be incompatible with GTR. One can obtain only that amount of mass-energy which was known in advance to exist inside a given surface. The possibility of reformulating

22. Two reservations must be made. First, we are considering GTR to be valid. Second, negative masses and regions with a negative energy density are presumed not to exist.

the steady-state theory of the Universe by replacing the creation of matter from the C-field by this inflow is thereby dashed.

We shall now discuss the "white-hole" hypothesis. The name is presently used to refer to "nuclei that are delayed during the cosmological expansion" (Novikov 1964). In particular, a white hole is a hypothetical body that expands from a singularity, eventually exploding outward from below its gravitational radius.

The possibility that such bodies exist was demonstrated by Novikov (1964*b*) and later independently by Ne'eman (1965). What is meant here is existence in the realm of classical GTR—without any account of quantum effects—in an expanding Universe. According to this hypothesis, white holes are nuclei of matter delayed in the cosmological expansion of a Friedmann model due to local inhomogeneities in the initial conditions. The expansion delay can be specified arbitrarily with respect to the time of an external observer. The duration of the delay as well as the content and the kinetic energy of the matter flying apart are specified as initial conditions at the singularity or as close to the singularity as is desired. Finally, a delayed nucleus is surrounded totally by matter of lesser density (a "vacuole") or, in the limiting case, by a vacuum.

It is of course tempting—but not necessary—to compare the white-hole hypothesis with a whole group of partially understood astrophysical phenomena—in particular, with quasars and explosions of galactic nuclei.

The white-hole idea differs from Jeans's idea. Thus, what is considered in the white-hole idea is space with a singly connected topology. It is taken into account that the Universe as a whole expands and evolves and that the expansion begins at a singularity. Inside a white hole, in the vacuum part of a vacuole, the singularity in the very simple, spherical symmetric case is of the Schwarzschild type. Outside the white hole, the Friedmann metric applies.

The singularity (which, again, encompasses the whole Universe) is everywhere spacelike—both outside the white hole and inside it. That is, a system of reference exists in which this singularity occurs everywhere simultaneously. Because of the inhomogeneity of the singularity, however, a delay occurs (with respect to an external observer's time) in the expansion of part of the matter. This part expands later as seen by the external observer. Moreover, the delay as seen by the external observer can be arbitrarily long because its duration depends on the choice of the size of the Schwarzschild singularity encompassing the delayed matter.

For matter with vanishing pressure, $P = 0$, Novikov (1964*b*) constructed an exact solution for the case when a spherical mass M is replaced by a white hole of the same mass in a homogeneous, isotropic Universe.

When the delayed matter has nonzero pressure $P \neq 0$, and in particular when $P = \epsilon/3$, mathematical difficulties arise, upon a consideration of the internal region of the vacuole. These are not theoretical difficulties, how-

ever; they do not evoke doubts concerning the existence of the white-hole type of solution. If, however, $P \neq 0$ in the entire surrounding Universe, then the external solution is perturbed. Actually, in a homogeneous Universe with $P \neq 0$ but without a white hole, the mass inside a selected comoving volume decreases because of the adiabatic decrease of energy over the course of the expansion (see chap. 1). In the case of a white hole, the central nucleus does not expand, and, in the view of an external observer, its mass does not decrease. In addition, there is in the solution a pressure discontinuity at the boundary between the unperturbed Universe and the vacuum Schwarzschild solution. The pressure discontinuity leads to a flow of plasma from the outside to the internal region. A white hole thus gives rise to plasma accretion for any relation between the proper mass of the hole and the mass excised from the cosmological solution by the presence of the hole. Up to this time, only rough estimates of the accretion effect exist; a complete picture of the phenomenon has not been constructed. It remains unexplained whether, by allowing for accretion, one can make use of the white-hole hypothesis for an explanation of various astrophysical phenomena.

Besides the questions concerning accretion, however, there exists another side of the matter. It is connected first of all with the quite highly anisotropic expansion of space that occurs in a white hole in the vacuole near the Schwarzschild singularity, in distinction to the isotropic expansion outside the white hole in the isotropic Universe. A second point is that given the white-hole hypothesis, the distant observer "sees" the Schwarzschild singularity for all of the time of the delay. This is due to the fact that below the gravitational radius r_g of a white hole is located the so-called T_+ region (see Vol. 1), in which all particles move exclusively away from the singularity toward r_g and indeed can emerge from below r_g toward an external observer.

As is shown in § 24.4 and § 24.5, the first circumstance, the anisotropic expansion near the singularity, leads to the intense creation of pairs. The second circumstance, that this takes place in the expanding T_+ region, gives rise to the possibility that the created particles are ejected from below r_g toward an external observer. This circumstance consequently can lead to a spontaneous decrease in the white hole's mass. In reality, as is shown in work done by Zel'dovich, Novikov, and Starobinsky (1974), the created particles very significantly change the spacetime metric below r_g. As a result, the created particles turn out to be "locked up" inside the white hole in the realistic case of a hot Universe. At first, during the very early stages of expansion of the Universe (for $t \ll r_g/c$), the particles do not emerge from below r_g because of the pressure of the surrounding gas in the hot Universe. Later (for t of the order of r_g/c and for $t > r_g/c$), because of self-gravitational closure, the gravity of the created particles does not permit them to move outward from below r_g. During the stage

when $t \gg r_g/c$, the white-hole mass decreases because of the aforementioned effect of adiabatic energy decrease in any comoving volume in the hot Universe. During a later stage, the white hole can lose only a small part of its mass due to outflow of created particles. (For numerical estimates of the effects of this process, see the study done by Zel'dovich, Novikov, and Starobinsky 1974.) It is quite essential, however, that all changes related to the created particles lead to an inhibition of the outburst of the delayed nucleus if the delay is to be much longer than r_g/c. This is because a white hole either explodes almost immediately after $t < r_g/c$, or never explodes (except in cases involving extremely degenerate initial conditions).

The basic conclusion is thus that in the hot Universe a white hole is observed as a mass of particles created near a Schwarzschild singularity. The matter these particles comprise expands and, in the view of a distant observer, the external boundary of the matter-filled region asymptotically approaches r_g from below. For times $t \gg r_g/c$, any radiation or outflow of matter from the white hole decays exponentially fast—notwithstanding that the approach of the matter to r_g proceeds in cadence with the expansion!

Thus, although a white hole can exist for an unlimited time, to an external observer it acquires the traits of a black hole, processes freeze, and the boundary of the matter approaches r_g. The qualitative difference, however, is that this object arises as a result of an explosion (quantum in nature) from below—from a singularity—to r_g, and not as a result of gravitational collapse, the fall toward r_g of originally rarefied material. Another process which transmutes a white hole into a black hole is described by Eardley (1974). He pointed out that even a small quantity of accreted matter hides a white hole under the common gravitational radius of a white hole and this matter.

Passing now from the physical nature of white holes to cosmology per se, we note the following. It is shown in previous works by the present authors (Zel'dovich and Novikov 1967*b;* Novikov and Thorne 1973) that if white holes exist, then, at the beginning of the cosmological expansion, the fraction of the mass included within white holes constitutes an insignificant fraction of the mass of all of the material. Actually, the matter constitutes an ultrarelativistic gas satisfying $P = \epsilon/3$ at the beginning of the expansion. In the course of the expansion, the energy of this gas in any chosen comoving volume decreases as a result of the adiabatic cooling. However, the mass contained in the white holes does not decrease after the white holes separate from the overall expansion (after the instant $t \approx r_g/c$).[23] In order that the total mass contained in white holes not be excessively large by the present time and not affect very substantially the expansion time of the Universe, it is necessary that the fraction of the total mass contained in them be sufficiently small at the start of the expansion. It was

23. We are supposing that the white holes do not possess relativistic peculiar velocities.

shown in the aforementioned works that this fraction α (which depends on time) must satisfy the following limit at early times:

$$\alpha < 10^5 z^{-1} \quad \text{for} \quad z > 10^5 .$$

We then obtain

$$\alpha < 10^{-3}$$

by the time t corresponding to r_g/c for a mass of the order of $10^9 M_\odot$, the order of a quasar mass. This is the time $t \approx 10^4$ s.

In conclusion, we emphasize that the subject of this section has been hypothetical objects—white holes. There are still no serious indications at all that such objects exist in the Universe. Indeed, arguments which at one time seemed to testify in favor of their existence, arguments based on the explosions of galactic nuclei and on quasars, have lost force in the course of time.

Returning to the question of local topology, we ought to mention an idea of Wheeler (1960). At the level of the very smallest scales—on the scale of the Planck length, $l_{\text{Pl}} = 10^{-33}$ cm—one can expect large metric fluctuations and a complicated topology. Wheeler has put forth a hypothesis concerning the structure of space that brings to mind the structure of plastic foam with its uncountable number of topological handles connecting the various nearby regions. For such scales it is difficult to confirm anything and difficult to refute anything. We know well, however, that the macroscopically averaged energy density (and, indeed, all components of T_{ik}) of this structure equals zero, or is negligibly small if the cosmological constant $\Lambda \neq 0$. Upon introducing the "plastic-foam idea," then, one must simultaneously also introduce such methods of renormalization as guarantee this macroscopic result. Observation and experiment—higher criteria of truth—commit us to this.

24.13 Statistical Physics and Gravitation

The statistical-physics description of systems in which gravity plays an important role has unusual, nontrivial, features. Some 30 or 50 years ago, investigators did not take sufficient account of these features. A superficial application of the principles of thermodynamics (i.e., of statistics) to cosmology then led to completely incorrect conclusions. So, in particular, there arose both the notorious problem of the "thermal death of the Universe" and as a consequence of incorrect assumptions, a wrong theory[24] of primeval fluctuations on cosmological length scales.

To understand the general features of the problem, we shall examine a fundamental quantity—the canonical partition function $Z = \int e^{-E/kT} dp dx$,

24. The connection between statistical fluctuations and the perturbations leading to formation of galaxies was discussed in Part 3.

which is an integral taken over the entire phase volume of the system.[25] The energy E of the system is the sum of the kinetic energy of the particles and the energy of their interaction, in particular of their gravitational interaction. In classical theory, the integral decouples into the product of two integrals, one over the momenta and the other over the coordinates. Hence, we must examine here the integral $\int e^{-U(x)/kT}dx$, where $U(x)$ is the potential energy of the gravitational interaction. The symbol x is used to represent all of the coordinates of all of the particles constituting the system. In Newtonian theory,

$$U(x) = -\frac{1}{2}G\sum_{i,k}(m_i m_k/|r_i - r_k|) .$$

With such an expression for the potential energy U, the statistical integral diverges. More precisely, the integral diverges when the particles become spread out so far apart (mean interparticle distance $r \to \infty$, $U \to 0$) that the integration volume per particle[26] becomes infinite. This property of the integral with gravitational potential energy does not differ from the analogous integrals for other types of interactions, those involving the electromagnetic and nuclear forces, for example. It is worth recalling that it is the divergence for large r that is important when considering a system of a finite number of particles in an infinite space—for example, a cluster of stars (see Vol. 1).

It is very important that the partition function for the gravitational interaction diverges for small $r \to 0$ also. This happens because the exponential increases more rapidly as $r \to 0$ than the volume decreases. The Coulomb electromagnetic interaction between two equally charged particles of opposite sign possesses the same property. However, quantum effects limit from below the minimum distance to which an electron can approach a nucleus. In a system of a large number of charged particles, moreover, the attraction between charges of opposite sign is compensated by the repulsion between charges of like sign. Coulomb forces are therefore responsible for the formation of atoms, fluids, and crystals, but do not lead to catastrophic effects.

Gravitational forces are meanwhile insignificantly small at the atomic level, but grow with an increase in the number of particles. On macroscopic, astronomical scales, the gravitational forces are such that the partition-function integral diverges at small interparticle distances, i.e., for large densities, so those forces then play a decisive role.

25. As is well known, the free energy of the system is expressed in terms of Z by $F = -kT \ln Z$. It is then easy to express all other quantities of interest. For example, the entropy S satisfies $S = -\partial F/\partial T$, the energy $E = F - T\partial F/\partial T$.

26. By itself the concept of an infinite volume and an infinite number of particles with finite number density gives no difficulty.

When noninteracting particles are considered, statistical physics leads to the conclusion that a uniform distribution of particles in space is the most probable type of distribution. In somewhat more scientific language, one can say that entropy is a maximum for a uniform, homogeneous matter distribution in a given volume with given total energy. By the second law of thermodynamics, the emergence of macroscopic motion or of magnetic fields is impossible in such a uniform distribution. Indeed, this is true regarding any organized structure.

In distinction to this, the presently observed state of the Universe is characterized by sharp inhomogeneity in the density and the temperature (as evidenced by stars and galaxies) and by macroscopic motions and magnetic fields.

The contrast between the observations and the equilibrium state suggests two questions:

1. In the future, must the Universe pass into a homogeneous, isothermal, equilibrium state (the thermal death)?

2. If the Universe was at some time in the past in a homogeneous, isothermal, equilibrium state, how did the presently observed, clearly nonequilibrium state arise? Should one ascribe its appearance to a gigantic fluctuation that violates the principles of thermodynamics?

The mistake in the formulation of these questions consisted of tacitly ignoring the role of gravitation in the analysis. With an account of gravitation, a homogeneous distribution does not correspond to the state of maximum entropy at all. To see this, consider gathering the matter together into isolated blobs—stars or galaxies. Given the conservation of the total energy, a release of gravitational energy thereupon takes place and the temperature rises.

Now the probability of any given state of the system is the product of the probability of the given spatial distribution of particles and the probability that the particles have the given velocity or kinetic energy distribution. Upon the gathering of the matter into blobs, the "space" probability decreases, but the "velocity" probability grows in compensation; on balance, the overall product grows. The formation of stars and galaxies from the uniformly distributed matter therefore occurs with entropy growth. It is a natural process, not requiring the violation of the laws of thermodynamics.

The evolution direction of stars, galaxies, and of the whole Universe is discussed in detail in many sections of this book. It would be inappropriate to repeat such a discussion here. In any case, the reader should understand clearly that entropy grows in such systems and that the evolution in the future, with a due regard for the role of gravity, by no means leads to a dull, homogeneous isothermal state for the Universe.

24.14 The Brans-Dicke Theory of Gravitation (BDT) and Its Cosmological Consequences

The[27] theory mentioned in the title (hereafter referred to as BDT for short; see Brans and Dicke 1961; Dicke 1962, 1964, 1968) is closely connected with Mach's principle. In this theory, as well as in GTR, a non Euclidean, curved, four-dimensional manifold—spacetime—is considered. Along with the known particles and fields, a still hypothetical, "previously unknown"[28] massless scalar field, the so-called ϕ field, is considered in this manifold. The authors introduce it by an unusual method, with the result that, in the theory of gravity and in cosmology, the ϕ field plays a role that differs substantially from the role played by other fields—for example, the scalar field describing neutral pions.

In the original formulation of BDT, the scalar curvature R is multiplied by the function ϕ in the Lagrangian, thereby changing the usual action of GTR, $\int R dV$. Thus the complete BDT action consists of the three terms:

$$S = \int \left[\frac{c^3}{16\pi G_0} (-\phi R + L_\phi) + \frac{1}{c} L_m \right] dV . \qquad (24.14.1)$$

Here, V is the four-dimensional volume with $dV = (\sqrt{-g}) d^4x^i$, and L_m is the Lagrangian of the ordinary, known particles and fields. The Lagrangian of the ϕ field, L_ϕ, also has a somewhat unusual form, namely,

$$L_\phi = \omega \frac{\phi_{,i}\phi^{,i}}{\phi} . \qquad (24.14.2)$$

The commas denote covariant partial differentation, and ω is a dimensionless numerical parameter. The specific properties of BDT are determined by the form in which the first term in equation (24.14.1) is written.

The main ideas behind such notation are that for $\phi = 0$ there is no theory of gravity, since the term describing the elasticity of space then disappears (compare these remarks with the discussion in § 2.5 of Vol. 1). On the other hand, the ϕ field turns out to depend on the distribution of material in the entire space. In BDT, therefore, in correspondence with Mach's ideas, only in the presence of (ordinary) material does it turn out to be possible to find a metric and to define a locally inertial coordinate system.

For the duration of this section, expressions of the type "it turns out well," "it turns out," and "it corresponds to" will continue to be used many times. We must say right away, however, that these expressions do not

27. L. E. Gurevich, A. M. Finkel'shteyn, and V. A. Ruban rendered us great assistance in the writing of this section. For this, we are sincerely grateful to them.

28. Here the terminology of the patent office is being parodied. In this case, however, the existence of a "previously unknown" field remains unproved and even doubtful at the present time.

mean that BDT is correct or that it describes nature adequately. From the point of view of Brans and Dicke, the "correspondence" with Mach's principle is a virtue of the theory. But Mach's principle has not been proved (see § 24.10), so a "correspondence" with it is not a proof of the correctness of BDT. The same assertion applies to other properties of BDT.

The differential equations of BDT are obtained, naturally, by a variation of the Lagrangian presented above. The variation of the metric leads to equations similar to the equations of standard GTR. The scalar field gives an ordinary contribution $T_{ik(\phi)}$ to the stress-energy tensor in addition to that of the ordinary material $T_{ik(m)}$. But the effective gravitational "constant" is no longer a universal constant. In its place enters the term const.ϕ^{-1}, with ϕ a function of the coordinates! We thus have

$$\frac{\phi c^4}{8\pi G_0}(R_{ik} - \tfrac{1}{2}g_{ik}R) = T_{ik(m)} + T_{ik(\phi)} ,$$

where G_0 is the constant just mentioned.

By varying the action with respect to the ϕ field, we obtain in addition the equation

$$2\omega\phi^{-1}\Box\phi - \frac{\omega}{\phi^2}(\phi^{,i}\phi_{,i}) + R = 0 . \tag{24.14.3}$$

But since R is in turn expressed in terms of $T_{(m)}$ and $T_{(\phi)}$, one can rewrite the equation in such a way as to make clear the role of the matter as the source of the field:

$$\Box\phi = \frac{8\pi G_0}{(3 + 2\omega)c^4} T_{(m)} . \tag{24.14.4}$$

The scalar $T_{(m)}$, the trace of the stress-energy tensor, enters on the right-hand side of the last equation. This trace is the energy density, or, more precisely, ρc^2 for dust (for which $P = 0$) or $\epsilon - 3P$ for a fluid. It is curious that $T_{(m)} \equiv 0$ for an ultrarelativistic gas and that consequently $\Box\phi = 0$ in this case. Note too that, as already mentioned, all of the operations indicated (differentiation and $\Box$) are carried out with respect to the non-Euclidean spacetime.

In Volume 1, various types of flat-spacetime gravitational theories were discussed: a tensor theory, a vector theory, and a scalar theory. In particular, the scalar theory mentioned there was connected with the names of Brans and Dicke (see § 2.7 in Vol. 1). More precisely, however, one ought to say that the role of the scalar function in BDT is more complicated than is the case in a flat-spacetime theory because the function ϕ is not analogous to a Newtonian scalar potential.

Also, BDT contains a dimensionless free parameter ω. Indeed, the observable consequences of the theory depend on the value chosen for ω; in the particular case $\omega = \infty$, BDT corresponds to standard GTR.

Note well that the primary characteristic feature of BDT is that the gravitational "constant" turns out to be a function of coordinates, $G \sim [\phi(x, y, z, t)]^{-1}$, and that this function obeys definite equations in which the matter present plays the role of source. It follows that the dimensionless quantities Gm^2/e^2 and $Gm^2/(\hbar c)$, where m and e are the masses and charges of elementary particles, are not constant. All of the remaining "physics," however—those phenomena characterized by the numbers $e^2/(\hbar c)$ and m_p/m_e—turns out to be the same over all space and for all time. In the sense that it is a theory in which G is not constant, BDT brings to mind the Jordan (1948) and Dirac (1938) theories. But a valuable distinction between BDT and these others (for which reason we are discussing BDT in detail) is that no change in the number of particles is presumed by it. Thus BDT does not come into conflict with the laws of laboratory physics.

As Dicke (1962) has shown, other, equivalent, formulations of the theory are also possible in which the ϕ field interacts directly with the scalar T describing the matter and in which the gravitational constant (without quotation marks) is constant. To make up for this, the effective masses m' of all particles depend on ϕ and consequently are not constant. The locally observable quantity $Gm^2/(\hbar c) = G(m')^2/(\hbar c)$ is then also not constant, as is also true in the first variant of BDT (see above).

At first glance such a variant of the theory is closer to the usual equations, and the results lend themselves to interpretation more easily. In reality, however, if the masses of the elementary particles are variable, then atomic clocks keep time at a nonconstant pace as a consequence. In this theory, therefore, the atomic unit of time and the atomic unit of length depend on the instantaneous local value of the scalar field ϕ, so time and length as measured with respect to atomic standards do not coincide with the time and length which enter into the definition of the metric $ds^2 = dt^2 - dl^2$ (after a local reduction to a Minkowski form). Complications therefore arise in the interpretation of any experiment. For this reason, we shall confine ourselves below to the first variant of the theory, the variant in which ϕ interacts directly with the curvature R, but not with the matter. In this case, the gravitational "constant" is not constant ($G \sim \phi^{-1}$), but, as compensation, the particle masses are constant.

In order to determine ϕ, not only are an equation and a source specification necessary, but so also are initial conditions and boundary conditions. Now distant sources influence the value of ϕ at a given point through a coefficient that is inversely proportional to the distance, i.e., that varies as r^{-1}; the number of such sources meanwhile grows with the distance in proportion to r^2dr. In the first approximation, therefore, ϕ is determined primarily by distant sources, so one can consider the ϕ field to be constant, independent of spatial coordinates. This would be so even on the scales of the Galaxy, the solar system, or the Earth, notwithstanding the gigantic

differences in density from the average cosmological value.[29] In a particularly simple cosmological solution, that for a homogeneous and isotropic flat model filled with "dust" (so $P = 0$) and having $\phi = 0$ when $t = 0$, the solution obtained is

$$\phi \sim t^s, \quad s = \frac{2}{4 + 3\omega}; \qquad a \sim t^q, \quad q = \frac{2 + 2\omega}{4 + 3\omega}. \tag{24.14.5}$$

In the limit $\omega \to \infty$, this solution reduces to the standard results: ϕ and G are constant, and $a \sim t^{2/3}$.

For $\omega = 6$ in particular, the change in the gravitational constant (or in the masses of particles) is predicted to be roughly

$$\frac{1}{G}\frac{dG}{dt} = -0.14H \approx -5 \times 10^{-12}\,\mathrm{yr}^{-1}, \tag{24.14.6}$$

where H is the Hubble parameter. In Dirac's (1938) theory, dG/dt is 10 times larger.

The accuracy of the known values of the interplanetary distance, now measured by means of radar, permits one to hope for a direct answer over the coming 10 year period to the question of whether there are such changes. Experiments conducted by Shapiro (1965) presently show that $G^{-1}dG/dt$ is not greater than $4 \times 10^{-11}\,\mathrm{yr}^{-1}$.

There are at the present time more interesting tests that do not require long-term experiments, though. Thus, in particular, the observable consequences of BDT have been developed in detail and show that BDT gives rise to corrections in the GTR predictions of the results of classical GTR tests. As an example, the angular deviation α of a ray passing near the Sun is expressed by the formula $\alpha = 1''.75[(3 + 2\omega)/(4 + 2\omega)] = 1''.75[1 - (4 + 2\omega)^{-1}]$ in BDT. The theory also changes GTR prediction of the secular shift in the perihelion of Mercury's orbit. Measurements of these effects in recent years have not been favorable from the standpoint of BDT. First of all, recent data give for α a value that leads to

$$|\omega| > 6 \quad \text{or} \quad |\omega| > 10.$$

Second, measurements of the shift in the perihelion of Mercury are, to a very high degree of accuracy, in agreement with the GTR predictions.

To save BDT, one of the authors, Dicke (1964), made the hypothesis that the gravitational field of the Sun possesses a significant quadrupole

29. Noticeable differences in ϕ from the average appear only where the gravitational potential of an isolated mass becomes of the order of c^2, where c is the speed of light. BDT therefore introduces quantitative changes into the theory of neutron stars and causes especially substantial qualitative changes in the theory of collapse and of black holes (Brans 1962 and Nariai 1969*a*, *b*). However, this point goes beyond the subject matter of this book.

moment. The coincidence between the measured shift in the location of the perihelion of Mercury's orbit and the GTR prediction of it is in Dicke's opinion a result of the fortuitous cancellation of two corrections one connected with BDT and the other dependent on the Sun's quadrupole moment. Dicke's demonstration that the Sun has a quadrupole moment was based on distortions in the shape of its photosphere, i.e., on flattening of the solar disk that he observed. New measurements have left the question open, however.

The consequences of BDT in the area of celestial mechanics are considered in several studies (Nutku 1969; Shapiro *et al.* 1971; Finkel'shteyn and Krainovich 1974). These consequences have, up to now, not found experimental confirmation.

BDT specifically predicts the existence of longitudinal gravitational waves or, more precisely, waves of the ϕ field. Measurements of the Weber type at the moment when a nearby supernova explodes could thus confirm BDT—or at least give a sensitive estimate of the parameter ω. Even for small deviations from GTR (for $|\omega| > 6$) the effect can be significant since, in distinction to the radiation of gravitational waves as described by GTR, the radiation of scalar waves also takes place even given spherically symmetric oscillations or collapse (Dicke 1964, Morganstern and Chiu 1967, Gurevich and Dynkin 1972). The "ordinary" gravitational waves predicted by GTR are tensor waves and therefore are radiated by an oscillating or a collapsing body only to the degree that the system deviates from spherical symmetry and has a quadrupole moment (see § 17.2 or, for more detail, Vol. 1). As is well known, however, a repetition of Weber's experiments has not confirmed the presence of the impulses; in addition, though, scalar waves have also not been observed.

Hence, an often encountered situation has arisen concerning BDT. There is no experimental proof that GTR is insufficient, and there are also no experimental arguments in favor of BDT. But thanks to the judiciously introduced parameter ω, all of the deficiencies associated with the rigid form of standard GTR are capable of being translated into a quantitative question. The results of the studies in which this viewpoint is presented are that BDT has not been refuted, but merely that BDT has undergone a quantitative limitation on its possible applicability.

What are the cosmological consequences of BDT? For a model filled with dust ($P = 0$) and given a definite choice of initial conditions, formulae (24.14.5) written above show how the exponent in the cosmological solution changes. The equation of state for $P = \epsilon/3$ yields a peculiar situation in BDT, however. Here just the deviations from this law, the difference $P - \epsilon/3$, is substantial for BDT.

Nucleosynthesis in BDT is considered in a study by Dicke (1968). In the first variant of the theory, the variant in which the masses of the elementary particles are constant and in which time measurements are

made with respect to atomic clocks, the essence of the change in nucleosynthesis reduces to an acceleration in the pace of the cosmological expansion. But we know that an accelerated expansion leads to an increase in the fraction of neutrons surviving after the hadron era. Given an increase in the rate of expansion by a factor of σ, with σ between 10 and 1000, the helium abundance turns out to be greater than 40%, which is inadmissible. For σ between 1000 and 10^5, the helium abundance is less than 10%, but the predicted amounts of deuterium (>3%) and ^{3}He ($> 2 \times 10^{-4}$) are then inadmissibly large. It is therefore necessary that $\sigma > 10^5$ hold. Specifically, Dicke advocates a scalar-tensor cosmological model with $\sigma = 1.5 \times 10^6$ (corresponding to $\omega = 5$) and with practically no helium abundance. If, however, $\omega > 6$ (as celestial mechanics and radar measurements in the solar system apparently indicate), then difficulties will arise in BDT cosmology!

Notice that in addition to "ordinary" solutions satisfying the initial conditions that $\phi = 0$ at $t = 0$, one can also consider in BDT solutions with $\phi \neq 0$ at $t = 0$ or with $\phi a^3 = \text{const.} \neq 0$, i.e., with a free ϕ field. These solutions possess curious features. Thus, the ϕ field behaves under definite conditions as maximally stiff ($P = \epsilon$) matter. Given a definite choice of the parameters $P = \epsilon < 0$, a smooth transition from contraction to expansion in an isotropic (Friedmann) model is also possible.

Greenstein (1968) (see also Morganstern 1973) has examined the case of positive ω, the case for which the singularity is preserved. Here, however, the singularity depends on the nature of the ϕ field and not on the nature of whatever matter is present. For studies of anisotropic solutions, see the papers by Belinsky and Khalatnikov (1972), Ruban and Finkel'shteyn (1973), and Matzner, Ryan, and Bion (1973).

Gurevich, Finkel'shteyn, and Ruban (1973) investigated the theoretically new case of negative ω. At first glance the absence of a singularity in this case realizes the ancient dream of theoreticians and of philosophers. Since the presence of a singularity signifies the necessity of enlisting new theoretical ideas, only a nonsingular solution is complete and logically closed. However, having eliminated the singularity, we find ourselves confronted with new questions: How does nucleosynthesis take place? In what way is a Planck equilibrium RR spectrum obtained? These questions arise because, though an infinite density is perhaps not necessary, a density of the order of 10^6–10^8 g cm^{-3} and a temperature higher than 10^{10} K are necessary for an understanding of the observable picture of the Universe.

But it is a theoretical difficulty that is the most important problem in the negative ω case. Getting rid of the singularity is possible only in the confines of a theory with a field that can give rise to a negative energy density. Without a negative energy density a singularity is inevitable—so say the theorems developed by Penrose and by Hawking and Penrose (see § 23.2). But a field giving a negative energy density is also capable of

giving rise to an "explosion of the vacuum." The quantum theory of such a field predicts that, in the vacuum (the standard, flat-space vacuum!), quanta of this field can be created along with ordinary particles. As Zel'dovich (1974) has shown, however, such a spontaneous creation of particles contradicts observations.

Let us now summarize our discussion of BDT. The diversity of solutions turns out to be significantly greater than the authors of the theory presumed. New variants not related to Mach's ideas have appeared, and the free ϕ fields have been considered. A significant number of publications concerned with BDT have appeared. In spite of this, general considerations advanced in recent years, as well as the observations, are unfavorable toward BDT. It therefore seems probable that interest in this theory will disappear in the near future. Just in 1976 important observations were published, which show no deviations from standard GTR to a level of 1%. In terms of BDT the authors give $\omega > 25$ with 90–99% confidence. The oblateness of the Sun is disproved. In effect, this is "the end of the chapter."

24.15 Cosmology and New Hypotheses in Field Theory

During the 1970s, a particular direction of research in quantum field theory and the theory of elementary particles experienced further development and became a popular subject for study—the idea of spontaneous symmetry violation or, for short, the idea of "broken symmetry." Briefly, the idea reduces to this: Definite symmetries are perceived in nature, but they turn out to be inexact.

An example of such a broken symmetry is the so-called CP-invariance, i.e., the symmetry with respect to mirror reflection (P = parity), accompanied by the replacement of particles by their antiparticles (C = charge conjugation).[30]

It is natural as a first approximation to construct a theory in which one symmetry or other is exact. It is well known, for example, just how to construct a Lagrangian (the expression from which, by variation, the equations governing the behavior of the particles and fields are obtained) so as to get equations with given symmetry properties, for instance, CP-invariance. In reality, though, it is well known that the symmetry is not exact: CP-invariance is indeed violated in several processes. Must we then

30. It was Landau (1958) who put forward the idea of a combined inversion, i.e., CP invariance. In everyday life, mirror reflection symmetry (parity invariance) holds to a high degree of accuracy even without charge conjugation. In beta-decay and in other weak-interaction phenomena, however, the violation of parity invariance reaches the 100% level: neutrinos are exclusively left-circularly polarized, while antineutrinos are exclusively right-circularly polarized. But these phenomena are invariant under combined inversion, i.e., CP-invariant. Indeed, at the present time, violations of CP invariance are observed only in the decay of neutral K mesons and are quite small (Christensen, Cronin, Fitch, and Turley 1964).

relinquish the use of an idealized Lagrangian? No. The idea of a broken symmetry preserves the symmetry of the Lagrangian, the symmetry of the idealized theory, but chooses such a functional dependence that an asymmetric solution is stable.

Let us explain this idea by means of an example. Suppose that the potential energy for some system is an even function of a coordinate: $U(x) = ax^2 + bx^4 + \ldots$. The curve of the potential energy is then symmetric with respect to the replacement of x by $-x$. Now the equilibrium states of a particle are obtained from the condition that $U(x)$ be a minimum at the particle location, i.e., from the condition $F = -\partial U/\partial x = 0$. Suppose further that $a > 0$ and $b > 0$ and that there are no other terms in the expression for U. The curve of the potential energy then has one minimum, at $x = 0$, since, as one ought to expect, a symmetric theory (one with $U[x] = U[-x]$) has a symmetric equilibrium state.

Now take instead the case with $b > 0$ but $a < 0$. The potential energy now has two minima, at $x_{1,2} = \pm(|a|/2b)^{1/2}$, where $U(x_{1,2}) = -|a|^2/(4b)$. The origin of coordinates, $x = 0$, at which $U = 0$, is a maximum of the curve. Hence, as before, the potential energy is symmetric, $U(x) = U(-x)$ but, in the equilibrium state, either one ($x_1 = \{|a|/|2b|\}^{1/2}$) or another ($x_2 = -\{|a|/[2b]\}^{1/2}$) asymmetric solution is realized.

One of the present authors (Z) vividly recalls how L. D. Landau, while studying the flow of a viscous fluid in an expanding pipe (a problem with symmetric equations and boundary conditions), noticed that the symmetric solution is unstable and that stable asymmetric flows exist in a definite interval of the Reynolds number. He recalls further what great theoretical significance Landau ascribed to this situation and how Landau rejoiced at the possibility of analytically following the solution beyond the point where the unexpected asymmetric solution arises.

N. N. Bogolyubov emphasizes in a similar vein that in statistical mechanics a destruction of isotropy, of the equivalence of all directions, takes place upon the formation of a ferromagnet or a crystal—even though the original equations treat all directions equivalently. From the point of view of mathematical physics, a broken symmetry in contemporary theories is considered a phenomenon of the same type but concerned with the simplest possible object—the vacuum.

Now when a scalar field ϕ is considered, it is assumed that the field energy contains, along with the usual ϕ^2 and $(\nabla\phi)^2$ terms, a term $U(\phi)$ in just the functional form that is written above, $U = a\phi^2 + b\phi^4 + C$, though with $a < 0$. The theory is of such a nature that $\phi = 0$ corresponds to complete symmetry. There is then no violation of CP-invariance. In reality, however, a vacuum state is realized for which U is a minimum—at the point where $\phi \equiv +(|a|/2b)^{1/2}$, for example. In this asymmetric state, the symmetry of other fields is also destroyed because of the connection between ϕ and the other fields (though we shall not write this connection

here). Thus, CP invariance is destroyed. The constant $C - |a|^2/(4b)$ is taken in such a way that the energy density of the asymmetric vacuum equals zero, in correspondence with the indications of astronomical data that the cosmological constant vanishes (or is very small). Consequently, the energy of the unstable symmetric state is positive: $U(\phi = 0) = C - |a|^2/(4b)$. Note finally that this account gives only general considerations concerning the broken-symmetry idea. It was not intended to be a rigorous discussion.

The idea of broken symmetry is also used in other theories. For example, one develops a theory where, at $\phi = 0$, electrons and muons have zero rest mass and therefore there is symmetry between e, μ, and neutrinos. The theory is constructed so that $\phi \neq 0$ is realized in the vacuum state, and it is the interaction of this ϕ with electrons and muons which gives them rest masses. Broken symmetry is also used to establish the symmetry between weak and electromagnetic interactions. Of course all these theories are not yet proved, but they are elegant and fashionable.

Kirzhnits (1973) made an important remark about the consequences of the broken-symmetry hypothesis as applied to the theory of the hot Universe. Given the conditions that hold near the singularity, when the temperature is high, one ought to expect the ϕ field (and the other fields too!) to be highly excited, in which case the average value of ϕ turns out to be zero notwithstanding that the energy density is not minimal. This is because the order disappears upon an increase in temperature. This effect is similar to the disappearance of ferromagnetism when a crystal melts. Other predictions of the symmetric theory (in which, for all practical purposes, $m_e = m_\mu = m_\nu = 0$ holds, and so forth) are also recovered when the high-temperature situation is examined. Linde (1974) suggests that one assume that the energy density of the vacuum—and hence the cosmological constant too—becomes large and positive in the high-temperature symmetric state.

It is customary to define the cosmological constant as the energy of the vacuum or, more precisely, as the limit of the energy density for a particle (excitation) density that tends to zero. For certain assumptions about the field Lagrangian the results are flawless. Caution is needed in connection with the use of the terminology "cosmological constant," however. As explained earlier (Vol. 1), the cosmological constant gives rise to an effective stress-energy tensor $T_i^k = \text{const.}\ \delta_i^k \Lambda$, giving $p_\Lambda = -\epsilon_\Lambda$, with a Pascalian isotropic pressure which is invariant under a Lorentz transformation and independent of coordinates in Minkowski space. But for a specified high temperature, one cannot take the particle density or, in particular, the density of field quanta to be zero. This is due to the fact that the particle density is specified by the temperature; and since the particles are at rest on the average only in one system of coordinates, the hot gas is not Lorentz invariant.

The energy density at a finite temperature has a definite value, but the separation of this density into two parts (into parts characteristic of matter and of vacuum) is an arbitrary procedure. Let us consider this in more detail using the ordinary definitions (the subscript "vac" means vacuum, while "m" means matter):

$$\epsilon = \epsilon_{\text{vac}} + \epsilon_m , \qquad P = P_{\text{vac}} + P_m .$$

We assume that $\epsilon_{\text{vac}} = -P_{\text{vac}} = \text{const.}$ and that $dE = -PdV$. The law of energy conservation is then valid for each part separately, so that

$$d(\epsilon V) = -PdV ; \qquad d(\epsilon_{\text{vac}} V) = \epsilon_{\text{vac}} dV = -P_{\text{vac}} dV ,$$

$$d(\epsilon_m V) = -P_m dV .$$

According to Linde's definition, i.e., for $\epsilon_{\text{vac}} \neq \text{const.}$, these equalities do not exist, so it is necessary to take account separately of the interaction of the vacuum with the matter. Although matter is not Lorentz invariant, one can extract a part of the total T_i^k, that of the interaction of the vacuum with the hot matter m, which is Lorentz invariant. In such a situation, one can divide ϵ into two parts, but it is not necessary, for only the total energy density ϵ and the total pressure P enter into any of the equations.

In a study by Zel'dovich, Kobzarev, and Okun' (1974), an analysis is carried out of the evolution of the hot cosmological model. In this analysis, the field equations used admit a symmetric solution for ϕ at high temperatures and a broken symmetry at low temperatures, following Linde. In one simple variant of the theory, the conditions for a broken symmetry are achieved in the course of expansion and cooling, and a mosaic structure (or, as physicists say, a domain structure) is obtained. The space disintegrates into a large number of regions, in some of which $\phi = \phi_+ \equiv +[|a|/(2b)]^{1/2}$ and in others of which $\phi = \phi_- \equiv -[|a|/(2b)]^{1/2}$. Correspondingly, the sign describing the violation of CP-invariance is different in the two types of regions. In particular, the sign for the difference between the probabilities for the two channels

$$K_L^0 \to \pi^+ + \mu^- + \bar{\nu}_\mu , \qquad K_L^0 \to \pi^- + \mu^+ + \nu_\mu ,$$

of the K_L^0 decay is characterized by the type of region.

Between the ϕ_+ and the ϕ_- region, a transition layer is formed that has a definite high surface energy, a surface tension, and a surface mass density. The high density and the tension of the domain walls allow one to think of the entire system as a gas satisfying the equation of state $P = -2\epsilon/3$ if the distribution of domains is random and if the motion of the domain walls is slow. It then turns out that the law giving the overall expansion of the Universe changes substantially, becoming $a \sim t^2$.

The disintegration of the space into a mosaic of ϕ_+ and ϕ_- regions must give rise to noticeable inhomogeneity in the density and in all other quantities even if the initial, high-temperature state is strongly homogeneous. The

high energy density of the walls (of the order of 10^{76} ergs cm^{-3} or 10^{55} g cm^{-3} by a rough estimate) leads to the conclusion that if the walls were left intact to the present time, they would give rise to an inadmissible distortion of the RR isotropy. Curiously, RR once again gives the coup de grace that limits the flights of one's fancy. We are sure that no walls exist in all space up to the present-day particle horizon.

Does this mean that it is necessary to repudiate entirely a theory that destroys the symmetry? We shall mention two variants from which the contradictions are removed.

In the first, a variant with a complex ϕ field, $\langle|\phi|\rangle = 0$ at high temperatures, and $|\phi| = \phi_0 = \text{const.}$ at low temperatures. The different regions of the space are distinguished by the phase of the quantity $\phi = \phi_0 e^{i\alpha(x)}$, though not by its magnitude. Such a theory does not explain the violation of CP-invariance, but it does serve to explain the nonzero mass of the electron and of the muon. Furthermore, the phase $\alpha(x)$ can change smoothly: Instead of walls, vortices (lines where the phase changes rapidly) occur in this case. They are not prohibited by cosmology and may perhaps prove useful for elementary particle physics.

In the second variant devoid of contradictions, a variant with a real value $\phi = \pm\phi_0$ and with a sharp boundary, it is possible for the charge asymmetry of the matter (the baryon excess) to give rise to an asymmetry between the ϕ_+ and ϕ_- regions. For example, ϕ_+ regions might be formed predominantly, so that by the present time all of space forms a single ϕ_+ region, with no massive walls anywhere.

The hypotheses set forth above, as well as the last variants just discussed, are still very far from experimental confirmation by means of the techniques of elementary-particle physics currently employed in accelerator experiments, for example. Why then have we presented this hypothesis here? We have done so to show how indefinite are our ideas concerning the physics of processes near the singularity and how numerous are the unexpected possibilities in this region.

24.16 An Oscillating Universe?

The idea of a static, unchanging Universe doubtless does not conform to reality and has been abandoned. Also not in correspondence with reality is the idea of a steady-state Universe (see § 24.9), in which the Hubble expansion is compensated by the creation of matter. However, the prolonged existence and popularity of these theories is startling! Recall that Einstein himself consciously altered GTR in such a way that the equations turned out to be compatible with a static cosmological solution (see chap. 4).

We have already remarked in § 3.12 that the idea of a steady state evidently has a definite intrinsic attractiveness. In the nineteenth century this attractiveness could perhaps have been connected with the apparent

stationarity of astronomical systems. In the twentieth century another viewpoint exists. According to it, the laws of electric, lepton, and—most importantly—baryon charge conservation, all of which hold absolutely in all known processes, are a weighty argument in favor of the eternal existence of the Universe, which possesses a nonzero baryon charge.

Cannot one rather consider the eternal existence of the Universe as stationary on the average by assuming that the evolution is oscillatory? In such a scheme, expansion follows the singular state and in turn smoothly slows and changes into contraction. Thereupon the contraction proceeds, becoming more rapid, and finally ends in collapse. It is assumed that the collapse of the Universe as a whole can change sign in the singular state and transform into anticollapse. In other words, collapse toward a singular state changes into expansion from a singular state. This change constitutes the beginning of a new cycle, whose behavior mimics that of preceding ones.

In order that expansion (experienced by us at the present time) change into contraction, it is necessary that the density in the Universe be greater than the critical density. The corresponding, supposed change in the radius of the Universe with time is shown in the upper part of Figure 67.

The question of the possibility of a collapse-anticollapse transition at the singularity (with quantum and other phenomena accounted for) remains open today. We shall be assuming here that such a transition is possible. Thus, the "bounce" in Figure 67, at which the curve approaches the abscissa and then is "reflected" from it, depicts the transition.

What are the consequences of such a theory? In the hot model of the Universe, there is one feature that is favorable with respect to the existence of an oscillating Universe; i.e., regardless of the chemical composition of the matter undergoing collapse, after the passage through the "fireball" singularity, the matter returns to the primordial composition—to the celebrated abundance of 70% H and 30% ^{4}He for isotropic expansion (see § 7.5).

It would seem that a repetition of cycles is possible. However, the second law of thermodynamics forbids an oscillating model. This is because the

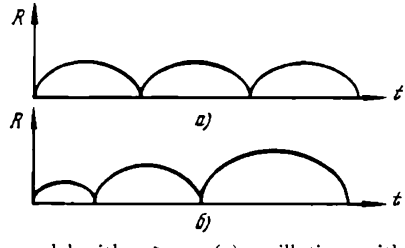


FIG. 67.—Oscillatory model with $\rho > \rho_c$: (*a*) oscillations without growth of entropy; (*b*) oscillations with growth of entropy.

entropy of the Universe only grows; indeed, it grows in the course of both expansion and contraction. Upon collapse, moreover, one can expect a particularly sharp increase in the entropy, for, at some instant during the late stages of contraction, the most general, "eight-function-dependent" type of cosmological solution, one with anisotropic contraction (see § 23.3), must be realized. Under these conditions, viscosity is important; the acceleration of neutrinos and the spontaneous creation of particles proceed particularly intensely. The entropy thereupon also grows significantly.

In drawing further conclusions, the assumption that the entropy does not decrease upon the passage though the singularity plays a central role. We shall make this assumption even though a consistent quantum theory of the singular state does not exist. The basis for such an extrapolation is one of the most important principles of modern theoretical physics—the correspondence principle. Thus, a future quantum-gravitational theory will contain within itself both GTR and the Newtonian theory of gravity as limiting theories—just as GTR itself contains Newtonian theory as a limiting case. Now in GTR and in Newtonian theory entropy only increases. A singular state probably need not lead to a violation of either this general law or the law of baryon conservation. But if entropy grows from one cycle to the next, then each succeeding cycle differs from the preceding one. Then, as Tolman (1934) showed long ago, a calculation leads to the existence of cycles isolated with respect to time, cycles that exhibit a growing amplitude or increasing maximum radius for the Universe (see the lower part of Fig. 67).

It is still not known precisely if our Universe is open or closed—if the density ρ is greater or less than the critical density ρ_c. Arguments in favor of an open Universe are somewhat more convincing, but one cannot consider a variant with $\rho > \rho_c$ as excluded. From this point of view, the oscillating variant of evolution is not excluded.

However, the specific entropy of our Universe, the entropy per baryon, is finite. It follows from this that the Universe has lived through only a finite number of cycles in the past and has a finite time of existence because in each cycle the entropy increases by a finite amount. For an infinite number of cycles, therefore, the specific entropy would be infinite; but this is not the case.

Thus, with due regard for the growth in the entropy, the oscillating model of the Universe does not allow one to describe a Universe with eternal existence from $t = -\infty$. The theory of the oscillating Universe does not achieve the aim intended for it, the aim of describing an eternal Universe.

From the point of view of an eternal Universe, the preferable picture is that of an open Universe, which contracts once in the past, during the time $-\infty < t < 0$, with the contraction turning into expansion at the $t = 0$ singularity. Unbounded expansion during the present stage, when $0 < t <$

t_0 (t_0 represents the present time), and in the future then follow in this picture. But, of course, since we can still not say whether this is indeed the true evolutionary picture of the Universe, let us return to the analysis of the oscillating Universe.

We shall illustrate by simple calculations the assertion made by Tolman concerning the effect of the entropy growth on the evolution of the cycles of expansion and contraction in a closed model. The calculation will be carried out for a homogeneous, isotropic cosmological model (see chap. 1). The metric for the model takes the form

$$ds^2 = c^2dt^2 - a^2(t)[dr^2 + \sin^2 r(d\theta^2 + \sin^2\theta d\phi^2)], \quad (24.16.1)$$

and the equation for the scale factor $a(t)$ is

$$\frac{1}{2}\left(\frac{da}{dt}\right)^2 = \frac{4\pi G}{3}\rho a^2 - \frac{c^2}{2}. \quad (24.16.2)$$

This homogeneous model is closed, so the law of nucleon conservation can be formulated in an elementary fashion; the number of nucleons N does not change from one cycle to another. Now the assumption was made concerning the entropy that it does not change at the moment of the jump from one cycle to another. However, it does increase over the course of each cycle. Indeed, the growth in the entropy is particularly large if stars form and nuclear reactions proceed during the course of a given cycle.

We shall now examine the energy density in a comoving coordinate system. For definiteness, we shall compare its values at the time in each cycle when the radius is a maximum, when the expansion changes into contraction. Thus, with the condition $da/dt = 0$, we find from equation (24.16.2) that

$$\epsilon = \rho c^2 = \frac{3c^4}{8\pi G a_{\max}{}^2} \quad 24.16.3)$$

at this time. But the energy density depends alternatively on the nucleon density and on the specific entropy, the entropy per baryon. These quantities are

$$n = \frac{\mathcal{N}}{2\pi^2 a^3}, \qquad s = \frac{S}{\mathcal{N}}, \quad (24.16.4)$$

respectively, where $\mathcal{N}$ is the total number of nucleons, S is the total entropy, and $2\pi^2a^3$ is the volume of the closed Universe. Now the relation

$$\epsilon(n, s) \equiv \epsilon\left(\frac{\mathcal{N}}{2\pi^2a^3}, \frac{S}{\mathcal{N}}\right) \quad (24.16.5)$$

is given by the equation of state of matter. With certainty one can say only that $\partial\epsilon/\partial s > 0$ and $\partial\epsilon/\partial n \geq \epsilon/n$, however. The equation

$$\epsilon\left(\frac{\mathcal{N}}{2\pi^2 a_{\max}{}^3}, \frac{S}{\mathcal{N}}\right) = \frac{3c^4}{8\pi G a_{\max}{}^2} \quad (24.16.6)$$

serves to determine the maximum radius in a given cycle in terms of the known entropy. It follows from the properties of ϵ that in each succeeding cycle, $a_{\max}$ is greater than in the preceding cycle (see Fig. 67).

We give here a formal proof of these assertions. We first introduce $E(n, s)$, the energy per baryon, which depends on n and s: $\epsilon = nE(n, s)$, where $dE = -PdV = -Pd(n^{-1}) = Pn^{-2}dn$ for $s = \text{const}$. From this it follows that $\partial E/\partial n > 0$, and $\partial\epsilon/\partial n = E + n(\partial E/\partial n) > E = \epsilon/n$ since $P > 0$. Hence $\partial(\ln \epsilon)/\partial(\ln n) > 1$. Now we write the equation describing the halt in the expansion in the form $a_{\max}{}^2\epsilon(n_{\max}, s) = 3c^4/(8\pi G) = \text{const}$. Here $n_{\max}$ is the density at the expansion maximum. We shall now find the logarithmic derivative of the left-hand side of this equation with respect to $a_{\max}$ and s. We obtain

$$\frac{\partial[\ln (a_{\max}{}^2\epsilon)]}{\partial[\ln a_{\max}]}\frac{da_{\max}}{a_{\max}} + \frac{\partial[\ln (a_{\max}{}^2\epsilon)]}{\partial s}ds = 0\,,$$

or

$$\left[2 + \frac{\partial(\ln \epsilon)}{\partial(\ln n)}\frac{d(\ln n)}{d(\ln a_{\max})}\right]\frac{da_{\max}}{a_{\max}} + \frac{\partial\epsilon}{\partial s}\frac{ds}{\epsilon} = 0\,.$$

Since $n = N/(2\pi^2a^3)$, $d(\ln n)/d(\ln a_{\max}) = -3$; and since $\partial(\ln \epsilon)/\partial(\ln n) > 1$, $2 - 3[\partial(\ln \epsilon)/\partial(\ln n)] = -k < 0$. It follows from this that

$$\frac{da_{\max}}{ds} = \frac{a_{\max}}{k\epsilon}\frac{d\epsilon}{ds} > 0\,.$$

With the growth in the entropy, the radius at the time of the maximum expansion grows. In short, a cyclic closed Universe expands upon heating.

The energy per baryon E is also greater at the time of the maximum expansion (when $da/dt = 0$) in each succeeding cycle, where

$$E = \frac{\epsilon}{n} = \text{const.} \cdot a_{\max}{}^3\epsilon = \text{const.} \cdot a_{\max}\,. \qquad (24.16.7)$$

As a consequence of gravitational interaction, the increase in the energy per nucleon does not contradict the law of energy conservation. The total energy of the closed Universe is always identically equal to zero.

In the oscillating model, therefore, the amplitude of the oscillations is always growing. A calculation shows that the change in the amplitude in each cycle depends on the change in the entropy, i.e., on the specific processes that occur. Hence, with due regard for gravitation and GTR, the irreversible growth in the entropy leads to a picture that is quite different from the picture of a thermal death (a Universe with a constant temperature everywhere and therefore quiescent). The latter is the picture that was sketched by nineteenth-century physicists (for more detail about this, see § 24.13).

Let us now extend the solution into the past. The minimum amplitude of a cycle in the past is determined by the relation

$$a = \left(\frac{3c^4}{8\pi G\epsilon_{\text{baryon}}}\right)^{1/2}, \qquad (24.16.8)$$

where $\epsilon_{\text{baryon}} \sim m_{\text{baryon}}c^2$. In other words, the minimum amplitude is determined by the magnitude of the masses of the nucleons. The amplitude is therefore always finite; but in each cycle the entropy changes by a finite amount. Since the entropy is now finite, only a finite number of cycles is possible in the past. The assumption that cycles are possible thus only pushes aside the difficulty concerning the existence of an initial singular state; it does not remove the difficulty.

New vistas are open now in connection with the formation of PBH—primordial black holes—and their Hawking-type evaporation. These possibilities are not yet explored. We shall here merely give some literature (Ruzmaykina and Ruzmaykin 1969; Nariai 1974) on the possibility of reversing the collapse by means of adding new terms to the Lagrangian or taking account of quantum vacuum polarization and particle formation.

24.17 The Creation of Gravitons near the Singularity

Gravitons appear in theory as an inevitable consequence of GTR, which predicts the existence of gravitational waves, and quantum mechanics, which requires the quantization of these waves. Classical theory, in this case GTR, further predicts that the speed of gravitational waves equals the speed of light, and consequently that the graviton rest mass is zero. In this sense gravitons are similar to photons (the quanta of the electromagnetic field). As was demonstrated by Grishchuk (1974), however, the equations describing gravitational waves are not conformally invariant with respect to a transformation of that average, large-scale metric on which background a gravitational wave is considered as a small, short-wave perturbation. This is in distinction to the situation for photons as described by Maxwell's equations. It is necessary to emphasize in this connection that the speed of a wave being equal to that of light and a vanishing rest mass for the associated particles are necessary but not sufficient conditions for the conformal invariance of the corresponding field.

In the scalar-field example, Penrose (1964), Tagirov and Chernikov (1968) have shown that, by generalizing the equations written for flat Minkowski space to equations for curved space, one can at will obtain conformally invariant or conformally noninvariant equations. But in the equations for gravitational waves, there is no arbitrariness because these equations are direct consequences of the GTR equations.

It turns out that the equations describing gravitational waves are not conformally invariant. In order to convince ourselves of this we shall examine a plane gravitational wave on the background of an expanding flat Universe (see § 12.5). The metric is thus

$$ds^2 = a^2(\eta)[d\eta^2 - \delta_{\alpha\beta}dx^\alpha dx^\beta] + h_{\alpha\beta}dx^\alpha dx^\beta . \qquad (24.17.1)$$

The conditions

$$h_\alpha{}^\beta = \gamma_\alpha{}^\beta e^{in\cdot r}\nu(\eta)\,, \qquad \gamma_\alpha{}^\beta n_\beta = 0\,, \qquad n = \text{const}\,. \tag{24.17.2}$$

then lead to the equation

$$\frac{d^2\nu}{d\eta^2} + 2\,\frac{1}{a}\,\frac{da}{d\eta}\,\frac{d\nu}{d\eta} + k^2\nu = 0 \tag{24.17.3}$$

for the wave amplitude. For a high frequency $k \gg a^{-1}da/d\eta$, the semiclassical approximation

$$\nu = \frac{Ae^{-ik\eta}}{a(\eta)} \tag{24.17.4}$$

holds. In this case, then, as was already noted (see chap. 17), the energy density of the gravitational waves changes as a^{-4}:

$$\epsilon = \nu^2\omega^2 = \nu^2k^2a^{-2} = \text{const.}\ a^{-4}\,. \tag{24.17.5}$$

The condition for the conservation of the number of gravitons per unit of comoving volume gives the number density of gravitons in physical space as $n \sim a^{-3}$. Since the wavelength is proportional to a, the frequency satisfies $\omega \sim a^{-1}$, so the energy density, in agreement with equation (24.17.5), satisfies

$$\epsilon = n\hbar\omega \sim a^{-4}\,.$$

The semiclassical approximation (24.17.4) corresponds to a situation wherein gravitons are neither created nor annihilated. But the semiclassical solution is an approximation. And it is just the abandonment of this approximation that results in the equations being not conformally invariant. This is so because the transformation from the Minkowski metric, with $a = \text{const.}$, to an evolving Universe, with $a \sim \eta$ or $a \sim \eta^2$, say, are examples of a conformal transformation. Let us now examine the situation in more detail.

In a manner similar to what was done in the theory of particle creation (Zel'dovich and Starobinsky 1971; see also § 24.5), we seek a solution of equation (24.17.3) in the form of a superposition of two oppositely moving waves:

$$\nu = \frac{f}{a}\,e^{-ik\eta} + \frac{g}{a}\,e^{ik\eta}\,. \tag{24.17.6}$$

By substituting this form for ν into equation (24.17.3) along with (in accordance with the method of Lagrange) the supplementary condition

$$\frac{d\nu}{d\eta} = \left(-ik - \frac{1}{a}\,\frac{da}{d\eta}\right)\frac{f}{a}\,e^{-ik\eta} + \left(ik - \frac{1}{a}\,\frac{da}{d\eta}\right)\frac{g}{a}\,e^{ik\eta}\,, \tag{24.17.7}$$

we obtain a first-order equation for the quantity f. It is

$$\frac{df}{d\eta} = \frac{i}{2ka}\,\frac{d^2a}{d\eta^2}\,(f + ge^{2ik\eta})\,. \tag{24.17.8}$$

An analogous equation that holds for g follows by making the replacements $f \to g$ and $i \to -i$ in this equation. It is apparent from this equation that f and g are constant and are not related to each other in the limit ($k \to \infty$) of short waves. For $k\eta \sim 1$, however, f and g are not constant. An initially pure f wave ($f = 1$, $g = 0$) gives rise to an oppositely moving g wave and is itself intensified.

In the quantum description, such a situation corresponds to the creation of a pair of gravitons in an isotropically expanding cosmological model. Recall that for a conformally invariant scalar field (see § 24.5), on the other hand, the creation of particle pairs takes place only when the metric is anisotropic,

$$ds^2 = dt^2 - a^2dx^2 - b^2dy^2 - c^2dz^2 . \qquad (24.17.9)$$

The creation then depends on differences of the type

$$[d(\ln a)/dt - d(\ln b)/dt]^2$$

so that it vanishes for isotropic expansion.

This important qualitative point is due to Grishchuk (1974). He notes that the connection between f and g is proportional to $d^2a/d\eta^2$. Consequently, the connection disappears not only when $a = \text{const.}$, but also when $a = \text{const.} \cdot \eta$. Indeed, such an expansion law does hold if the Universe is filled with matter for which $P = \epsilon/3$. Then

$$a \sim t^{1/2}, \quad \eta = \int \frac{dt}{a(t)} \sim t^{1/2}, \quad a \sim \eta , \qquad (24.17.10)$$

so that $d^2a/d\eta^2 = 0$. But if $P = \epsilon/3$, the trace of the stress-energy tensor satisfies $T = \epsilon - 3P = 0$. The scalar curvature R then vanishes even though neither R_{00} nor some components of R_{iklm} vanish. It is thus quite probable that the absence of graviton creation when $R \equiv 0$ is not accidental.[31]

From this analysis Grishchuk draws the conclusion that with an account of the back effect of the created particles on the cosmological expansion, the creation of gravitons brings the equation of state to the equation $P = \epsilon/3$ even if the matter does not have such an equation of state from the very beginning. If the primordial Universe is initially filled with matter for which $P_m \ll \epsilon_m$, then the expansion takes place with $d^2a/d\eta^2 \neq 0$, and gravitons are created until the time when the condition $\epsilon_\mu \sim \epsilon_m$ is achieved and the equation of state becomes $P = \epsilon/3$. In the course of the subsequent expansion, the role of the gravitons decreases; it turns out to be negligible at the present time.

If bouncing from collapse $a \sim \sqrt{-t}$, $t < 0$ to expansion, $a > \sqrt{t}$, $t > 0$ occurs (see § 24.16 on the oscillating Universe), the very moment of bounce is singular. It gives birth to a definite spectrum of gravitational waves,

31. Curiously, scalar particles are not created in the conformally flat metric for any variant of the wave equation when $R = 0$; see Parker (1968, 1969).

independent of the details of the bounce (notwithstanding that no creation occurs before and after the bounce, due to $R = 0$): see Starobinsky (1976).

Let us suppose, however, that the initial state is such that the matter is maximally stiff and obeys the equation of state $P_m = \epsilon_m$. In this case, $T \neq 0$ and $R \neq 0$, so graviton creation occurs. If the graviton creation turns on at $t_{\text{Pl}} \sim 10^{-43}$ s, and then essentially ends at $2t_{\text{Pl}}$ then the density of the created gravitons turns out to be of the order of the matter density. Subsequently when the graviton creation is essentially turned off, their density falls as a^{-4}, whereas the density of the "ordinary," though maximally stiff, matter falls as a^{-6}, given the stiff equation of state. According to the data of present-day observations, a Universe would thereupon be obtained with an inadmissible predominance of gravitons.

By way of summation, one can state that the creation of gravitons takes place in an isotropic cosmological model for which $P \neq \epsilon/3$ and is an argument against the existence of maximally stiff matter obeying the equation of state $P = \epsilon$ near the singularity (see § 24.8).

24.18 The Singularity and Conformal Invariance

Contemporary investigators (see, for example, Penrose 1968) often make use of conformal transformations to study singularities. We shall now outline this technique.

Suppose that with the help of a conformal transformation one can successfully associate a given, singular, four-dimensional manifold—the given, or "prototype," Universe—with another, nonsingular manifold—the "standard" Universe. One can generally study the propagation of waves in the standard Universe without theoretical difficulties. Upon so doing, one can then determine the nature of wave propagation in the prototype by means of the laws of the conformal transformation. Especially simple is that case when the standard model is a flat Universe with a Minkowski metric, a situation that necessitates that the prototype be conformally flat. Conformal flatness in turn means that the Weyl curvature tensor equals zero, but the Riemann curvature tensor R_{iklm} need not equal zero.

It is understood that the waves whose propagation is being considered obey conformally invariant wave equations. Electromagnetic waves in particular (but not gravitational waves; see the previous section) possess these properties.

An example of a conformally flat Universe is the Friedmann Universe. Near the singularity, its spatial curvature is not important[32] for any physical processes, so we write the metric in the following form:

$$\begin{aligned} ds^2 &= dt^2 - a^2(t)(dx^2 + dy^2 + dz^2) \\ &= a^2(\eta)(d\eta^2 - dx^2 - dy^2 - dz^2) \,. \end{aligned} \tag{24.18.1}$$

32. For a conformally flat expression for the open Universe, see Fock (1961).

The latter equality demonstrates that the standard Universe for the Friedmann model is the Minkowski model.

For a redshift $z < 3 \times 10^8$, the Universe is filled principally with photons and neutrinos, i.e., with particles of zero rest mass that are described by conformally invariant equations. Considering ever larger values of z, we recall that positron-electron pairs appear first, and hadrons (including baryons and antibaryons) thereafter, all of which have a nonzero rest mass m. However, upon the approach to the singularity ($z \to \infty$, $t \to 0$), the ratio of the rest mass m to the average particle energy becomes smaller and smaller. For particles (fields) with $m \neq 0$, there is no conformal invariance, though near the singularity one can speak of an approximate, asymptotic ($t \to 0$) conformal invariance.

In this approximation, Hoyle and Narlikar (1963) consider as a prototype a Universe filled with all forms of particles—photons, neutrinos, and hadrons—near a singularity. For $t \to 0$, these particles in aggregate form constitute an ultrarelativistic plasma with a temperature tending to infinity upon the approach to the singularity. The standard, obtained by means of a conformal transformation, is Minkowski space filled with a relativistic plasma. But the plasma is composed not only of truly massless particles—photons and neutrinos—but also of massless electrons and positrons, massless protons, and so forth. Thus, what is under discussion is fictitious, nonexistent particles, all of whose properties—charge, spin, and interaction—coincide with the properties of the true electron or proton, for example, with the exception that $m = 0$ for each fictitious particle. The temperature of the relativistic plasma can be selected arbitrarily, for here there is no quantity mc^2 with which to compare kT.

The electromagnetic, strong, and weak interactions of the particles do not destroy the conformal invariance of the plasma as a whole. Its equation of state is $P = \epsilon/3$ in the standard as well as in the prototype (near the singularity in the latter).

In the unchanging Minkowski model, the standard, constant-temperature stationary plasma passes through the surface $\eta = 0$ in the direction from $\eta < 0$ to $\eta > 0$ without experiencing any kind of peculiar occurrence.

Hoyle and Narlikar suppose that by returning from the standard to the prototype, one can obtain a description of the transition from contraction to expansion—from the situation at $t < 0$ to that at $t > 0$—in a Friedmann cosmological model. On considering such a possibility, we write the equation connecting t and η in the prototype and note first of all the following: In a model filled with relativistic plasma, the relations $a \sim t^{1/2}$, $d\eta = dt/a(t)$, $\eta \sim t^{1/2}$, and $t = \text{const.} \cdot \eta^2$ are fulfilled. But it follows from this that both regions of the standard, that for which $\eta < 0$ and that for which $\eta > 0$, correspond to the same sign for the true time t. Thus, the problem posed, to describe the passage through the singularity, is not solved in this case by enlisting the aid of the notion of conformal invariance.

Another remark concerning the Hoyle and Narlikar hypothesis is that even for massless particles, the gravitational interaction is not conformally invariant, just as the equations for gravitons are not conformally invariant.[33] There is a subtle point in the interpretation of the gravitational interaction: the collective gravitational interaction of the particles is taken into account in the evolution law $a \sim t^{1/2}$. But the average gravitational potential and the average metric do not completely describe the interaction. In the equation of state, it remains to take account of the so-called exchange interaction, which is a correction related to the correlation that exists between the locations of identical particles. In Sakharov's (1966) opinion, upon consideration of this interaction, a maximum temperature then arises that sharply changes the equation of state. The effects become of the order of unity when the temperature is of the order of the "Planck temperature," which satisfies $kT_{\text{Pl}} = m_{\text{Pl}}c^2$, so $T_{\text{Pl}} = 10^{32}$ K.

Hence, although for a different reason, the Planck time reckoned from the singularity again turns out to be that characteristic barrier beyond which one cannot manage to penetrate without utilizing new ideas. The idea of conformal invariance, valid for all fields (as $t \to 0$) except the gravitational field itself, permits one to advance further in studying the physics of the interval from 10^{-43} s to 10^{-15} s. To surmount the barrier at 10^{-43} s, however, this idea is insufficient.

Now, in light of the notion of conformal invariance, we shall consider the connection between the gravitational field and other fields. The basic idea of Einstein's theory of gravity is that space and time comprise a four-dimensional curved manifold (see § 1.2 in Vol. 1). Free particles move in this manifold along geodesic lines. Proceeding from this notion, one can find the trajectories of particles in the curved spacetime and the action of the spacetime curvature on the fields and particles (the electromagnetic field, electrons and positrons, neutrinos, and so forth). This action includes the creation of particles in a space whose metric depends on time and the vacuum polarization, the appearance of nonzero momentum and energy densities that depend on the curvature.

Another side of the theory of gravitation consists of ascribing a definite "elasticity" to spacetime itself. The qualitative idea goes back to Clifford. If space can in principle be curved, then there must be a factor—the elasticity—by virtue of which space, with high precision, actually remains flat on the scales of the Earth, the solar system, and the Galaxy. Indeed, the elasticity idea corresponds to the term const.$\int R dV$ in the GTR Lagrangian introduced by Einstein. Upon variation, this particular term gives the left-hand side of Einstein's equations, while the right-hand side of the equations represents the energy and momentum densities of real matter.

33. This is probably related to the particular dimensionality of the gravitational constant. From this constant G, together with $\hbar$ and c, one can construct a quantity with the dimensions of mass.

In Volume 1, an idea developed by Sakharov (1967b) is set forth. According to it, the elasticity of space depends on vacuum polarization. From this point of view, the theory of gravity cannot even in principle exist without an account of all other fields (the electromagnetic, positron-electron, and neutrino fields, etc.), although the total practical contribution of these fields is expressed through the magnitude of the gravitational constant G.

The above account gives the situation as it was in 1971 (see Vol. 1 and also Misner, Thorne, and Wheeler's recent book [1973]). New developments that radically change the picture are allowed when one takes into account the conformal invariance of the fields describing the γ, $e^{\pm}$, ν, . . . , and presumably the hadrons too. Indeed, it follows from the conformal invariance that the contribution of the enumerated fields and particles to the elasticity of the vacuum is negligibly small, of the order of $Gm^2/(\hbar c)$ or $[Gm^2/(\hbar c)]$ ln $[p_0/(mc)]$, where p_0 is a limiting momentum. Particles with zero rest mass contribute nothing at all; such a conclusion follows immediately from the realization that there is no vacuum polarization if $m = 0$ in a conformally flat space, a space with $ds^2 = a(x, y, z, t)(dt^2 - dx^2 - dy^2 - dz^2)$, but possibly with $R_{iklm} \neq 0$. It is, however, necessary to remove gravitons (see § 24.17) from the list of particles because the equation describing gravitational waves and their quanta, gravitons, is not conformally invariant.

In principle, therefore, a theory is possible in which only the spacetime is considered and in which interesting results are obtained concerning the gravitons and the curvature of the background spacetime due to the gravitational fields of the gravitons. The gravitons, in addition, propagate in this background spacetime, so what is obtained is actually a closed picture for a graviton-background system; the inclusion of the other particles and fields is unnecessary. Such a theory has not yet been constructed, however.

The starting point for such a theory is the set of commutation relations for the quantities describing the metric. Given the ordinary normalization of the gravitational waves made by introducing the field amplitude r, so that $\epsilon \sim r^2$, these relations take the symbolic form

$$[r_\alpha{}^\beta, r_\alpha{}^\gamma] = \left[\frac{1}{\sqrt{G}} h_\alpha{}^\beta(x) , \frac{1}{\sqrt{G}} \dot{h}_\delta{}^\gamma(x')\right] = \delta_\alpha{}^\gamma \delta_\delta{}^\beta \delta(x - x')\hbar .$$

As a result of taking account of these commutation relations for short-wave metric perturbations, a contribution to the Lagrangian of the form $S = (16\pi G)^{-1} \int R dV$, must (or in any case, can) be obtained. Now in quantum theory, the quantity $\psi = e^{iS/\hbar}$ enters; i.e., the quantity

$$\frac{i}{16\pi G\hbar} \int R dV$$

constitutes the exponent.

Thus, the product Gh (though not G or h separately or to other powers) enters into the commutation relations for the metric, and this same product[34] enters into the expression for the wave function of the entire system. Note that Gh is just equal to c^3 times the square of the Planck length. The internal logical closure of the construction under consideration is thereby confirmed. Here, closure of the theory does not of course exclude the possibility of considering other particles and fields in the theory (which would then require new components). What has been verified is merely that the propagation of gravitational waves and the overall smooth spatial curvature to which the waves give rise can be completely understood and described by including quantum effects in a theory with two quantities having dimensionality: the Planck length l_{Pl} and the speed of light c. The inclusion of other particles and fields is unnecessary, although their existence is not denied.

At the present time, persistent and successful attempts to unify the electromagnetic and weak interactions have been undertaken. In prospect are attempts to unify them with the theory of hadrons. An explanation of the theoretical questions concerning the connections between these various fields is therefore particularly important.

24.19 The Direction of Time

Without any basis at all, the physicist of the eighteenth century presumed that there exists both an absolute space and a spatially independent time having a specified direction of flow from the past to the future. He could then solve, for example, the mechanical problem of motion both for a determination of the future (predictions of eclipses) and for a determination of the past (calculations of past eclipses). He would do the latter by arranging the solution to agree with the known state at a given instant of time according to the laws of mechanics.

The special theory of relativity, which connects time and space by means of the Lorentz transformations, changed nothing regarding the question of the direction of time. As is well known, light cones (the loci of points $r = ct$ and $r = -ct$), which serve to delimit the "absolute future" and the "absolute past" with respect to a chosen "event" (the four-dimensional point $r = 0$, $t = 0$), are invariant under a Lorentz transformation. According to the special theory of relativity, therefore, a trip into the past by means of a Wells time machine is as impossible now as it must have seemed in the time of Newton.

The question of the direction of time arose about 100 years ago in connection with the theorems of statistics and of thermodynamics that are

34. It is evidently $l_{\mathrm{Pl}}^2 = 10^{-66}\,\mathrm{cm}^{-2}$; the speed of light is everywhere taken to be unity here. In the recent work of t'Hooft and Veltman (1974) the correction of the type $G^{-1}\int R dV$ cancels by invariant renormalization.

concerned with the irreversible growth of entropy in a closed system. There were proposals to define the future as that direction of time for which growth in the entropy takes place. Comparatively recently, Gold (1962) offered an idea concerning a possible connection between the direction of time (the "arrow of time") and neutrinos. In several studies, including a paper delivered by Hoyle in Moscow (in November of 1963), it is suggested that one define the positive direction of time with respect to the expansion of the Universe, i.e., with respect to the increase in the distance between galaxies. If galaxies were moving closer together at some moment in the past, then time too, according to Hoyle, flows at that moment in the past in the direction opposite to which it flows today.

In our opinion, attempts to connect the direction of time with some particular complicated phenomena or other are theoretically incorrect. A distinction between the past and the future exists in any process, including one encompassing a system of two particles. Physics is local, so phenomena taking place that involve only a single pair of particles must not depend on the growth in the entropy of some other complex system not interacting with the particles in question, or on the recession of galaxies.

The reversibility of the laws of mechanics and electrodynamics—the possibility of replacing t by $-t$ in the equations with no change in the physics—is often advanced as an objection to the last assertion. To solve a given problem, however, it is necessary to specify initial conditions in addition to giving a general solution of the equations. In a theory in which a field is being considered (for definiteness, the electromagnetic field), we, moreover, specify a *radiation condition* which is asymmetric with respect to the past and the future. Charges in fact do radiate electromagnetic waves when they move, with the energy of the charges themselves decreasing as time passes. And the amount of energy lost is proportional to the elapsed time during which the charges radiate. Because of this, one can think of the energy loss as resulting from the action of the wave field on the very charges that give rise to the field.

By virtue of the time reversibility of the equations, one can specify a field configuration consisting of converging waves in such a way that the waves' action on the charges would increase the energy of the system. It is clear, however, that such a formulation of the problem is artificial. There is no reason for the incident external waves to have just the frequency and phase necessary to drive the system. On the other hand, the waves emitted by the system are found in phase with the motion of the system by virtue of their origin. Namely, this phase relation guarantees that the field of an emitted wave is such as to slow the motion of the charges and take energy away from them. Thus, the radiation condition, or, to use a more general form of expression, the principle of causality, permits one to define unambiguously the future and the past.

Attempts to eliminate field concepts from theory by replacing them with

action-at-a-distance notions and by considering on an equal basis the retarded and advanced interactions were undertaken 30 years ago in connection with difficulties in elementary-particle theory (Wheeler and Feynman 1945). In such a theory, each particle interacts with the remaining particles, but not with itself, so as to avoid an infinity in its self-energy. Soon after this theory was developed, however, a correct method for examining the interaction of a particle with that field to which the particle itself gives rise (the so-called renormalization theory) led to results in brilliant agreement with precise experiments—with measurements of the Lamb-Rutherford shift and the anomalous magnetic moment of the electron. The theories involving action at a distance were thereby refuted (see Feynman 1948, 1967).

Hence the direction of time as specified in the mathematical language of physics by the radiation condition exists objectively in nature.

The irreversibility of thermodynamic phenomena is a consequence of the laws that apply to the elementary particles and to the elementary phenomena of particle collisions. Thus, for example, the law describing the growth of entropy is formulated only after the notion of future is itself defined. From this (inasmuch as the law $ds/dt > 0$ has already been established) it follows that it is convenient to make use of entropy growth as a practical means for determining the direction of increasing time. However, thermodynamics does not bear on the theoretical question of the determination of the direction of the arrow of time.

Finally, there is no foundation at all behind attempts at establishing a connection between the arrow of time and cosmology, i.e., with the expansion of the Universe. It is true that we are living in an expanding Universe, or at least in a Universe which, at the present time, is expanding. But can one raise this particular property to the rank of a principle and make use of it as a definition of the direction of the arrow of time? Can one say that the "future is the epoch when the radius of the Universe (or, equivalently, the distance between a given pair of distant galaxies) is greater than at the present time"? Such a view, as was already mentioned above, has indeed been expressed in the literature. But in adopting this point of view, we would also have to adopt its consequences. Thus, in a closed Universe (one for which $\Omega > 1$), contraction begins after the maximum radius $a_{\max}$ is achieved. Hence, if the arrow of time is determined cosmologically, then, at the time corresponding to $a = a_{\max}$, the arrow of time changes direction.

But this conclusion is hardly convincing. To see this, imagine a projectile launched from the Earth at a speed v_0, less than the escape velocity from the Earth: $v_0 < 11$ km s^{-1}. Such a projectile at first ascends, getting farther from the Earth, and then, upon achieving a definite maximum altitude $h_{\max}$, begins to fall. Clearly, no reformulation of any of the laws of physics takes place in the projectile at the altitude $h_{\max}$. In particular, a top (a gyroscope) located in the projectile moves monotonically. A transition from

expansion to contraction in a closed Universe is completely analogous to the transition from ascent to descent for the projectile. It is therefore entirely clear the arrow of time does not change direction at the instant of maximum expansion of the Universe. If it were to change direction, then rays of light in a contracting Universe, for example, instead of being emitted by stars and proceeding into the cosmos, would be incident onto the stars from which they were emitted. But this is obviously nonsense. Indeed, if $\Omega > 1$, then after the change from expansion to contraction, the density of radiation in the Universe is still very small for a very long time. Stars would continue to radiate light, and all local processes in the Universe would similarly continue to take place in the same temporal direction. The connection between the arrow of time and the expansion is a (very important, of course) property of our Universe at the present time. But this connection cannot be used as a definition of the arrow of time—as a definition of the concept of future.

In short, the only excuse we have for mentioning these erroneous views about the arrow of time here is that they have appeared so persistently in the literature.

REFERENCES

Abell, G. O. 1958, *Ap. J. Suppl.*, **3**, 211.

———. 1961, *Astr. J.*, **66**, 607.

———. 1962, in *Proc. IAU Symposium 15, Problems of Extragalactic Research*, ed. G. McVittie (New York: Macmillan), p. 213.

———. 1965, *Ann. Rev. Astr. Ap.*, **3**, 1.

Adamyants, R. A., Alekseyev, A. D., and Kolosnitsin, N. I. 1972, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **15**, 277.

Agekyan, T. A. 1970, *Zvezdi, Galaktiki, Metagalaktika* ("Stars, Galaxies, and the Universe") (Moscow: Nauka).

Akhiezer, A. I., and Berestetsky, V. B. 1965, *Kvantovaya Elektrodinamika* (Moscow: Nauka) (English translation: *Quantum Electrodynamics* [New York: Wiley-Interscience]).

Alexander, J. K., Brown, L. W., Clark, T. A., Stone, R. G., and Weber, P. R. 1969, *Ap. J.* (*Letters*), **157**, L163.

Alfvén, H. 1965, *Rev. Mod. Phys.*, **37**, 652.

———. 1971, *Phys. Today*, **24**, 28.

Alfvén, H., and Elvius, A. in *Proc. IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).

Alfvén, H., and Klein, O. 1962, *Zs. f. Phys.*, **23**, 187.

Alpher, R., and Herman, R. 1953, *Ann. Rev. Nucl. Sci.*, **2**, 1.

———. 1975, *Proc. Am. Phil. Soc.*, **119**, 325.

Ambartsumyan, V. A. 1960, *Nauchniye Trudy* ("Collected Works"), Volumes 1 and 2 (Yerevan: Publishing House of the Academy of Sciences of the Armenian SSR).

Ambartsumyan, V. A., and Kazyutinski, V. V. 1973, *Voprosy Filosofii*, **3**, 91.

Arons, J., and McCray, R. 1969, *Ap. J.*, **158**, 291.

Arp, H., and Bertola, F. 1971, *Ap. J.*, **163**, 195.

Bagrov, V. G., Bozrikov, P. V., Gitman, D. I. 1973, *Izv. Vuzov., ser. Radiofiz.*, **16**, 129.

Bahcall, J. N., and Hills, R. E. 1973, *Ap. J.*, **179**, 699.

Bardeen, J. M., and Wagoner, R. V. 1971, *Ap. J.*, **167**, 359.

Barger, V., Whisnant, K., Cline, D., and Phillips, R. J. H. 1980, preprints COO-881-135 and COO-881-146.

References were translated and edited by JoAnn Boyd Anderson, of Caltech.

Bechi, C., Gallinaro, G., and Morpurgo, G. 1965, *Nuovo Cimento*, **39**, 409.
Beckman, J. E., Ade, P. A. R., Huizinga, J. S., Robson, E. I., Vickers, D. G., and Harries, J. E. 1972, *Nature*, **237**, 154.
Bekenstein, J. D. 1972, *Phys. Rev.*, **D7**, 2333.
Belinsky, V. A., and Khalatnikov, I. M. 1969*a*, *Zh. Eksp. Teoret. Fiz.*, **56**, 1700.
———. 1969*b*, *Zh. Eksp. Teoret. Fiz.*, **57**, 2163.
———. 1972, *Zh. Eksp. Teoret. Fiz.*, **63**, 121.
Belinsky, V. A., Lifshitz, E. M., and Khalatnikov, I. M. 1970, *Uspekhi Fiz. Nauk*, **102**, 463.
———. 1971, *Zh. Eksp. Teoret. Fiz.*, **60**, 1969.
———. 1972, *Zh. Eksp. Teoret. Fiz.*, **62**, 1606.
Berestetsky, V. B. 1976, *Uspekhi Fiz. Nauk*, Vol. **120**, No. 3.
Berestetsky, V. B., Lifshitz, E. M., and Pitaevsky, L. P. 1968, *Relyativistskaya Kvantovaya Teoriya* (Moscow:Nauka) (English translation: *Relativistic Quantum Theory* [Oxford: Pergamon, 1971]).
Berezinsky, V. S. 1970, *Yadern. Fiz.*, **11**, 399.
Berezinsky, V. S., and Zatsepin, G. T. 1969, *Phys. Letters*, **28B**, 423.
Berger, B. K. 1972, University of Maryland, Tech. Report 73-024.
———. 1974, *Ann. Phys.*, **83**, 458.
Bergkvist, K. E. 1969, *Topical Conference on Weak Interactions* (Geneva: CERN), Vol. **69-7**, p. 91.
———. 1971, *Reviews of Particle Physics, Rev. Mod. Phys.*, **43**, No. 2, 11.
———. 1972, *Nucl. Phys.*, **B39**, 317.
Bertotti, B. 1966, *Proc. Roy. Soc. London*, **294**, 195.
Bianchi, L. 1918, *Lezioni sulla teoria dei gruppi continui finiti di trasformazioni*. Pisa: Spoerri.
Biermann, L. 1950, *Zs. Naturforsch.*, **5a**, 65.
Bisnovatyi-Kogan, G. S. 1968, *Astr. Zh.*, **45**, 74.
———. 1971, *Astrofizika*, **7**, 121 and 223.
———. 1972*a*, *Astr. Zh.*, **49**, 1238.
———. 1972*b*, *Zh. Eksp. Teoret. Fiz.*, **62**, 1593.
Bisnovatyi-Kogan, G. S., Kulikov, Yu. N., and Chechetkin, V. M. 1976, *Astr. Zh.*, **53**, 975.
Bisnovatyi-Kogan, G. S., and Mikhailovskiy, A. B. 1973, *Astr. Zh.*, **50**, 312.
Bisnovatyi-Kogan, G. S., Ruzmaykin, A. A., and Sunyaev, R. A. 1973, *Astr. Zh.*, **50**, 210.
Bisnovatyi-Kogan, G. S., and Sunyaev, R. A. 1971, *Astr. Zh.*, **48**, 881.
Bisnovatyi-Kogan, G. S., and Vainshtein, S. I. 1971, *Ap. Letters*, **8**, 151.
Bisnovatyi-Kogan, G. S., and Zel'dovich, Ya. B. 1970, *Astr. Zh.*, **47**, 942.
Blair, A. G., Beery, J. G., Edeskuty, F., Hiebert, R. D., Shipley, J. P., and Williamson, K. D. 1971, *Phys. Rev. Letters*, **27**, 1154.
Boccaletti, D., DeSabbata, V., Fortini, P., and Gualdi, G. 1970, *Nuovo Cimento*, **70B**, 129.

Bogdanova, L. M., and Shapiro, I. S. 1974, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **20**, 217.
Bogolyubov, N. N., Struminsky, B. V., and Tavkhelidze, A. N. 1965, Preprint O.I.Ya.I., D-1968.
Bogoyavlensky, O. I., and Novikov, S. P. 1973, *Zh. Eksp. Teoret. Fiz.*, **64**, 1475.
Böhm-Vitense, E., and Szkody, P. 1973, *Ap. J.*, **184**, 211.
Bondi, H. 1961, *Cosmology* (2d ed.; London and New York: Cambridge University Press).
Bondi, H., and Gold, T. 1948, *M.N.R.A.S.*, **108**, 252.
Bonnor, W. B. 1957, *M.N.R.A.S.*, **117**, 104.
Bonometto, S. A., DeZotti, G., and Lucchin, S. 1977, *Astr. Ap.*, **54**, 405.
Bortolot, V. J., Clauser, J. F., and Thaddeus, P. 1969, *Phys. Rev. Letters*, **22**, 307.
Bowyer, C. S., Field, G. B., and Mach, J. E. 1968, *Nature*, **217**, 32.
Boynton, P. E., and Partridge, R. B. 1973, *Ap. J.*, **181**, 243.
Boynton, P. E., and Stokes, R. A. 1974, *Nature*, **247**, 528.
Boynton, P. E., Stokes, R. A., and Wilkinson, D. T. 1968, *Phys. Rev. Letters*, **21**, 462.
Brachmachary, R. L. 1965, *Nuovo Cimento*, **2**, 850.
Braginsky, V. B. 1966, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **3**, 69.
Braginsky, V. B., Grishchuk, L. P., Doroshkevich, A. G., Zel'dovich, Ya. B., Novikov, I. D., and Sazhin, M. V. 1973, *Zh. Eksp. Teoret. Fiz.*, **65**, 1729.
Braginsky, V. B., Manukin, A. B., Popov, E. I., Rudenko, V. N., and Khorev, A. A. 1974, *Zh. Eksp. Teoret. Fiz.*, **66**, 801.
Braginsky, V. B., Zel'dovich, Ya. B., Martinov, V. K., and Migulin, V. V. 1967, *Zh. Eksp. Teoret. Fiz.*, **52**, 29.
———. 1968, *Zh. Eksp. Teoret. Fiz.*, **54**, 120.
Brans, C. 1962, *Phys. Rev.*, **125**, 2194.
Brans, C., and Dicke, R. 1961, *Phys. Rev.*, **124**, 965.
Bridle, A. N. 1967, *M.N.R.A.S.*, **136**, 219.
———. 1968, *Nature*, **219**, 1136.
Bronshtein, M. P. 1934, *Phys. Z. Sowiet.*, **2**, 100.
Brown, F. A. 1964, *M.N.R.A.S.*, **127**, 517.
———. 1968, *M.N.R.A.S.*, **138**, 527.
Bunner, A. N., Coleman, P. C., McCammon, D., Kraushaar, W. L., Palmieri, T. M., Shilepsky, A., and Ulmer, M. 1969, *Nature*, **223**, 1222.
Burbidge, E. M., and Sargent, W. L. W. 1971, in *Nuclei of Galaxies*, ed. D. J. K. O'Connell (Amsterdam: North-Holland).
Burbidge, G. R. 1971, *Nature*, **233**, 36.
Burbidge, G. R., and O'Dell, S. L. 1972, *Ap. J.*, **178**, 583.
———. 1973, *Ap. J.*, **183**, 759.
Byram, E. T., Chubb, T. A., and Friedman, H. 1970, *Science*, **169**, 366.

Callan, C. G., Dicke, R. H., and Peebles, P. J. E. 1965, *Amer. J. Phys.*, **33**, 105.
Carpenter, R. L., Gulkis, S., and Sato, H. 1973, *Ap. J.* (*Letters*), **182**, L61.
Carr, B. J. 1975, *Ap. J.*, **201**, 1.
———. 1976, *Ap. J.*, **206**, 8.
Chapline, G. F. 1976, *Nature*, **261**, 550.
Chernikov, N. A., and Shavokhina, N. S. 1973, *T.M.F.*, **16**, 77.
Chernin, A. D. 1969, *Astrofizika*, **5**, 656.
———. 1970, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **11**, 317.
———. 1971, *Ap. Letters*, **8**, 31.
———. 1972, *Ap. Letters*, **10**, 125.
Chernin, A. D., and Eidel'man, E. D. 1969, *Astrofizika*, **5**, 654.
———. 1971, *Astrofizika*, **7**, 314.
Chibisov, G. V. 1972*a*, *Astr. Zh.*, **49**, 74.
———. 1972*b*, *Astr. Zh.*, **49**, 286.
———. 1972*c*, doctoral dissertation, Moscow Physical-Technical Institute.
———. 1976, *Pis'ma Astr. Zh.*, **2**, 131.
Chincarini, G., and Rood, H. J. 1975, *Nature*, **257**, 294.
Chiu, H. Y. 1961, *Ann. Phys.*, **16**, 321.
———. 1964, *Ann. Phys.*, **26**, 364.
———. 1966, *Phys. Rev. Letters*, **17**, 712.
Christensen, J. H., Cronin, J. W., Fitch, V. L., and Turley, R. 1964, *Phys. Rev. Letters*, **13**, 138.
Christodoulou, D. 1970, *Phys. Rev.*, **D25**, 1596.
Chupka, W. A., Schiffer, J. P., and Stevens, C. M. 1966, *Phys. Rev. Letters*, **17**, 60.
Collins, C. B., and Hawking, S. W. 1973*a*, *Ap. J.*, **180**, 317.
———. 1973*b*, *M.N.R.A.S.*, **162**, 307.
Cowsik, R., and McClelland, J. 1972, *Phys. Rev. Letters*, **29**, 669.
Dashevsky, V. M., and Slysh, V. I. 1965, *Astr. Zh.*, **42**, 863.
Dashevsky, V. M., and Zel'dovich, Ya. B. 1964, *Astr. Zh.*, **41**, 1071.
Dautcourt, G. 1969, *M.N.R.A.S.*, **144**, 255.
Davies, P. C. W. 1974, *Nature*, **249**, 208.
Davis, M., Groth, E. J., and Peebles, P. J. E. 1977, *Ap. J.* (*Letters*), **212**, L107.
de Vaucouleurs, G. 1969, *Ap. Letters*, **4**, 17.
DeWitt, B. S. 1967*a*, *Phys. Rev.*, **160**, 113.
———. 1967*b*, *Phys. Rev.*, **162**, 1195.
DeWitt, C., and Wheeler, J. A. 1968, *Battelle Rencontres* (New York: W. A. Benjamin).
Dicke, R. H. 1962, *Phys. Rev.*, **126**, 1875.
———. 1964, in *Gravitation and Relativity*, ed. H. Y. Chiu and W. F. Hoffman (New York: W. A. Benjamin).
———. 1968, *Ap. J.*, **152**, 1.

———. 1970, *Gravitation and the Universe*, Jayne Lectures for 1969 (Philadelphia: American Philosophical Soc.).
Dicke, R. H., and Peebles, P. J. E. 1968, *Ap. J.*, **154**, 838.
Dicke, R. H., Peebles, P. J. E., Roll, P. G., and Wilkinson, D. T. 1965, *Ap. J.*, **142**, 414.
Dirac, P. A. M. 1937, *Nature*, **139**, 323.
———. 1938, *Proc. Roy. Soc.*, *A*, **165**, 199.
———. 1974, *Proc. Roy. Soc.*, *A*, **338**, 439.
Dolgov, A. D., Zakharov, B. U., and Okun', L. B. 1973, *Phys. Letters*, **46B**, 90.
Dolgov, A. D., and Zel'dovich, Ya. B. 1981, *Rev. Mod. Phys.*, in press.
Doroshkevich, A. G. 1965, *Astrofizika*, **1**, 255.
———. 1966, *Astrofizika*, **2**, 37.
———. 1967, *Astrofizika*, **3**, 175.
———. 1968, doctoral dissertation, Institute of Applied Mathematics, Moscow.
———. 1970, *Astrofizika*, **6**, 581.
———. 1973, *Ap. Letters*, **14**, 11.
Doroshkevich, A. G., Longair, M. S., and Zel'dovich, Ya. B. 1970, *M.N. R.A.S.*, **147**, 139.
Doroshkevich, A. G., Lukash, V. N., and Novikov, I. D. 1971, *Zh. Eksp. Teoret. Fiz.*, **60**, 1201.
———. 1973, *Zh. Eksp. Teoret. Fiz.*, **64**, 1457.
———. 1974, *Astr. Zh.*, **51**, 940.
Doroshkevich, A. G., and Novikov, I. D. 1964, *Doklady Akad. Nauk SSSR*, **154**, 809.
———. 1970, *Astr. Zh.*, **47**, 948.
Doroshkevich, A. G., Ryaben'kii, V. S., and Shandarin, S. F. 1973, *Astrofizika*, **9**, 258.
Doroshkevich, A. G., and Shandarin, S. F. 1973, *Astrofizika*, **9**, 549.
———. 1974, *Astr. Zh.*, **51**, 41.
———. 1975, *Astr. Zh.*, **52**, 9.
Doroshkevich, A. G., and Sunyaev, R. A. 1969, *Astr. Zh.*, **46**, 955.
Doroshkevich, A. G., Sunyaev, R. A., and Zel'dovich, Ya. B. 1973, in *Proc. IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
Doroshkevich, A. G., and Zel'dovich, Ya. B. 1963, *Astr. Zh.*, **40**, 807
———. 1975, *Ap. Space Sci.*, **35**, 55.
Doroshkevich, A. G., Zel'dovich, Ya. B., and Novikov, I. D. 1967*a*, *Astr. Zh.*, **44**, 295.
———. 1967*b*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **5**, 119.
———. 1967*c*, *Zh. Eksp. Teoret. Fiz.*, **53**, 644.
———. 1967*d*, *Astr. Tsirk.*, No. 442.
———. 1968, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **8**, 95.

———. 1969*a*, *Problemy Teoreticheskoi Fiziki* (Problems in Theoretical Physics), Papers dedicated to N. N. Bogolyubov (Moscow: Nauka).
———. 1969*b*, *Astrofizika*, **5**, 15.
———. 1971, *Zh. Eksp. Teoret. Fiz.*, **60**, 3.
Dreicer, H. 1964, *Phys. Fluids*, **7**, 735.
Dyer, C. C., and Roeder, R. C. 1973, *Ap. J.* (*Letters*), **180**, L31.
Eardley, D. M. 1974, *Phys. Rev. Letters*, **33**, 442.
Efimov, N. V. 1961, *Vysshaya Geometriya* (Higher Geometry) (Moscow: Nauka).
Einasto, Ya., Chernin, A. D., Kaasik, A., Kalamis, P., and Vennik, Ya. 1974, preprint No. 3, Tartu.
Einasto, Ya., Haud, U., Jôeveer, M., and Kassik, A. 1976, *M.N.R.A.S.*, **177**, 357.
Einasto, Ya., Kaasik, A., and Saar, E. 1974, *Nature*, **250**, 309.
Einasto, Ya., Saar, E., Kaasik, A., and Chernin, A. D. 1974, preprint No. 2, Tartu.
Einasto, Ya., Saar, E., Kaasik, A., and Traat, R. 1974, *Astr. Tsirk.*, No. 811.
Einstein, A. 1966, *Sobr. Nauchnikh Trudov* (Collected Scientific Works) (Moscow: Nauka).
Einstein, A., and Strauss, E. 1945, *Rev. Mod. Phys.*, **17**, 120.
Eisenhart, L. P. 1947, *Neprerivniye Gruppy Preobrazovanii* (Continuous Groups of Transformations) (Leningrad: I. L.).
Ellis, G., and McCallum, M. A. N. 1969, *Comm. Math. Phys.*, **12**, 108.
Elsasser, W. M. 1950. *Rev. Mod. Phys.*, **22**, 1.
Epstein, R. I., Arnett, W. D., and Schramm, D. N. 1974, *Ap. J.* (*Letters*), **190**, L13.
Evans, D. S. 1972, (ed.), *Proc. IAU Symp. No. 49.*
Evans, N. J., II, Zuckerman, B., Morris, G., and Sato, T. 1975, *Ap. J.*, **196**, 433.
Ewing, M. S., Burke, B. F., and Staelin, D. H. 1967, *Phys. Rev. Letters*, **19**, 1251.
Faber, S., and Jackson, R. E. 1976, *Ap. J.*, **204**, 668.
Fadeev, L. D., and Popov, V. N. 1967, *Phys. Letters*, **258**, 29.
Felten, J. E., and Morrison, P. 1966, *Ap. J.*, **146**, 686.
Fermi, E., and Yang, C. N. 1949, *Phys. Rev.*, **76**, 1739.
Feynman, R. P. 1948, *Phys. Rev.*, **74**, 939.
———. 1963, *Acta Phys. Polon.*, **24**, 697.
———. 1967, *Uspekhi Fiz. Nauk*, **91**, 29.
Feynman, R. P., Leighton, R. B., and Sands, M. 1965, *The Feynman Lectures on Physics* (Reading, Mass.: Addison-Wesley).
Field, G. B. 1965, *Ap. J.*, **142**, 531.
———. 1973*a*, in *Stars and Stellar Systems*, Vol. **9**, *Galaxies and the Universe*, ed. A. and M. Sandage and J. Kristian (Chicago: University of Chicago Press).

1973*b*, in *Proc. I.A.U. Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
Field, G. B., and Hitchcock, J. L. 1966*a*, *Phys. Rev. Letters*, **16**, 817.
———. 1966*b*, *Ap. J.*, **146**, 1.
Field, G. B., and Perrenod, S. C. 1977, *Ap. J.*, **215**, 717.
Finkel'shteyn, A. M., and Krainovich, V. Ya. 1974, *Celestial Mechanics*.
Fitremann, M., and Frisch, U. 1969, *Compt. Rend.*, **B268**, 705.
Fock, V. A. 1961, *Teoriya Prostranstva, Vremeni i Tyagoteniya, Izd. 2-e* (Theory of Space, Time, and Gravitation, 2d ed.) (Moscow: Fizmatgiz).
Ford, L. H. 1975, *Phys. Rev.*, **D11**, 3370.
Fowler, W. A. 1970, *Comments Ap. Space Phys.*, **2**, 134.
———. 1972, in *Cosmology, Fusion and Other Matters*, ed. F. Reines (London: Adam Hilger Ltd.), p. 67.
Frautschi, S., Steigman, G., and Bahcall, J. N. 1972, *Ap. J.*, **175**, 307.
Freeman, K. C. 1970, *Ap. J.*, **160**, 811.
Fricke, J. 1973, *Ap. J.*, **189**, 535.
Friedman, A. A. 1963, *Uspekhi Fiz. Nauk*, **80**, 439, 447.
Friedmann, A. 1922, *Z. Phys.*, **10**, 377.
———. 1924, *Z. Phys.*, **21**, 326.
Fulling, S. A., and Parker, L. 1974, *Ann. Phys.*, **87**, 176.
Gamow, G. 1946, *Phys. Rev.*, **70**, 572.
———. 1948, *Phys. Rev.*, **74**, 505.
———. 1952, *Phys. Rev.*, **86**, 251.
———. 1954, *Proc. Nat. Acad. Sci.*, **40**, 480.
———. 1968, *Proc. Nat. Acad. Sci.*, **59**, No. 2, 313.
Gell-Mann, M. 1964, *Phys. Letters*, **8**, 214.
Gershtein, S. S., and Zel'dovich, Ya. B. 1966, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **4**, 174.
Gertsenshteyn, M. E. 1961, *Zh. Eksp. Teoret. Fiz.*, **41**, 113.
Ginzburg, V. L. 1971, *Uspekhi Fiz. Nauk*, **103**, 87.
———. 1975, *Pis'ma. Zh. Eksp. Teoret. Fiz.*, **22**, 514.
Ginzburg, V. L., Kirzhnits, D. A., and Lyubushin, A. A. 1971, *Zh. Eksp. Teoret. Fiz.*, **60**, 451.
Ginzburg, V. L., and Ozernoy, L. M. 1965, *Astr. Zh.*, **42**, 943.
Ginzburg, V. L., and Syrovatsky, S. I. 1963, *Proiskhozhdenie Kosmicheskikh Luchei* (The Propagation of Cosmic Rays) (Moscow: Fizmatgiz).
Gold, T. 1962, in *Recent Developments in General Relativity* (Oxford, New York: Pergamon Press; New York: MacMillan), p. 225.
Goldberg, V. Z., and Chechetkin, V. M. 1976, preprint, Institute of Applied Mathematics, Moscow.
Goldsmith, D. W., Habing, H. J., and Field, G. B. 1969, *Ap. J.*, **158**, 173.
Goldsmith, D. W., and Silk, J. 1972, *Ap. J.*, **172**, 563.
Golonetsky, S. V., and Mazets, E. P. 1971, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **14**, 201.

Goloncts'ky, S. V., Mazets, E. P., Il'inskii, V. N., Aptekar', R. L., Bredov, M. M., Bur'yan, Ya. A., and Panov, V. N. 1971, *Ap. Letters*, **9**, 69.
Gott, J. R., and Rees, M. J. 1975, *Astr. Ap.*, **45**, 365.
Gould, R. J., and Ramsay, W. 1966, *Ap. J.*, **144**, 587.
Greenstein, G. 1968, *Ap. Space Sci.*, **2**, 155.
Gregory, S. A., and Tifft, W. G. 1976, *Ap. J.*, **205**, 716.
Grib, A. A. 1974, *Priroda*, No. 4.
Grib, A. A., and Mamaev, S. G. 1969, *Vadern. Fiz.*, **10**, 1276.
———. 1971, *Vadern. Fiz.*, **14**, 800.
Grishchuk, L. P. 1967*a*, doctoral dissertation, Sternberg Astronomical Institute, Moscow University.
———. 1967*b*, *Zh. Eksp. Teoret. Fiz.*, **53**, 1699.
———. 1967*c*, *Astr. Zh.*, **44**, 1017.
———. 1970, *Doklady Akad. Nauk SSSR*, **190**, 1066.
———. 1971, *Bulletin de L'Academie Polonaise des Sciences.* **XIX**, 12.
———. 1973, in *Proc. IAU Symposium 33, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
———. 1974, *Zh. Eksp. Teoret. Fiz.*, **67**, 825.
Grishchuk, L. P., Doroshkevich, A. G., and Lukash, V. N. 1971, *Zh. Eksp. Teoret. Fiz.*, **61**, 3.
Grishchuk, L. P., Doroshkevich, A. G., and Novikov, I. D. 1968, *Zh. Eksp. Teoret. Fiz.*, **55**, 228.
Grishchuk, L. P., Doroshkevich, A. G., and Udin, V. M. 1972, *Proceedings of the Third Soviet Gravitation Conference* (Yerevan).
Gull, S. F. 1976, Report to 16th IAU General Assembly.
Gull, S. F., and Northover, K. J. E. 1975, *M.N.R.A.S.*, **173**, 585.
Gunn, J. E. 1967, *Ap. J.*, **147**, 61.
Gunn, J. E., Lee, B. W., Lerche, I., Schramm, D. N., and Steigman, G. 1978, *Ap. J.*, **223**, 1015.
Gunn, J. E., and Peterson, B. A. 1965, *Ap. J.*, **142**, 1633.
Gurevich, L. E., and Dynkin, S. D. 1972, *Zh. Eksp. Teoret. Fiz.*, **63**, 369.
Gurevich, L. E., Finkel'shtein, A. M., and Ruban, V. A. 1973, *Ap. Space Sci.*, **22**, 231.
Guyot, M., and Zel'dovich, Ya. B. 1970, *Astr. Ap.*, **9**, 227.
Hagedorn, R. 1965, *Suppl. Nuovo Cimento*, **3**, 147.
———. 1969, Ref. Th. 1027-CERN.
———. 1973, *Cargese Lectures in Physics*, ed. E. Schatzman, p. 643.
Harkins, J. D. 1917, *J. Am. Chem. Soc.*, **39**, 856.
Harrison, B. K., Thorne, K. S., Wakano, M., and Wheeler, J. A. 1965, *Gravitation Theory and Gravitational Collapse* (Chicago: University of Chicago Press).
Harrison, E. R. 1968, *Ap. Letters*, **3**, 133.
———. 1970*a*, *Phys. Rev.*, **D1**, 2726.
———. 1970*b*, *M.N.R.A.S.*, **147**, 279.

Hauser, M. G., and Peebles, P. J. E. 1973, *Ap. J.*, **185**, 757.
Hawking, S. W. 1966*a*, *Ap. J.*, **145**, 544.
———. 1966*b*, *Proc. Roy. Soc. London*, *A*, **294**, 511.
———. 1974, *Nature*, **234**, 30.
———. 1975, *Comm. Math. Phys.*, **43**, 199.
Hawking, S. W., and Carr, B. 1973, in *Proc. IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
———. 1974, *M.N.R.A.S.*, **168**, 399.
Hawking, S. W., and Ellis, G. F. 1973, *The Large Scale Structure of Space-time* (Cambridge: Cambridge University Press).
Hawking, S. W., and Penrose, R. 1970, *Proc. Roy. Soc. London*, **A314**, 528.
Hawking, S. W., and Taylor, R. T. 1966, *Nature*, **209**, 1278.
Hayakawa, S. 1973, *Fizika Kosmicheskikh Luchei, Ch. I* (Physics of Cosmic Rays, Part I) (Moscow: Mir).
———. 1974, *Fizika Kosmicheskikh Luchei, Ch. II* (Physics of Cosmic Rays, Part II) (Moscow: Mir).
Hayashi, C. 1950, *Progr. Theor. Phys.*, **5**, 224.
Heckmann, O. 1942, *Theorien der Kosmologie* (Berlin).
Heckmann, O., and Schücking, E. 1955, *Zs. f. Ap.*, **38**, 95.
———. 1956, *Zs. f. Ap.*, **40**, 81.
———. 1959, in *Handbuch der Physik*, ed. S. Flügge (Berlin: Springer), p. 53.
———. 1962, Chapter 11 of *Gravitation: An Introduction to Current Research*, ed. L. Witten (New York: Wiley), p. 438.
Hegyi, D. I., Traub, W. A., and Carleton, N. P. 1974, *Ap. J.*, **190**, 543.
Henry, R. C., Fritz, G., Mecking, J. E., Friedmann, H., and Byram, E. T. 1968, *Ap. J.* (*Letters*), **153**, L199.
Hewish, A., Readhead, A. C. S., and Duffet-Smith, P. J. 1974, *Nature*, **252**, 657.
Hill, H. A., Clayton, P. D., Patz, D. L., Healy, A. W., Stebbins, R. T., Oleson, J. R., and Zanoni, C. A. 1974, *Phys. Rev. Letters*, **33**, 1497.
Holmberg, E. 1969, *Ark. f. Astr.*, **5**, 305.
Houck, J. R., and Harwit, M. O. 1969*a*, *Science*, **164**, 1271.
———. 1969*b*, *Ap. J.* (*Letters*), **157**, L45.
Howell, T. F., and Shakeshaft, J. R. 1966, *Nature*, **210**, 1318.
———. 1967, *Nature*, **216**, 753.
Hoyle, F. 1948, *M.N.R.A.S.*, **108**, 372.
———. 1958, in *La Structure et l'Evolution de l'Univers*, 11 Conseil de Physique Solvay (Brussels: Stoops).
———. 1969, *Nature*, **223**, 936.
Hoyle, F., and Fowler, W. A. 1973, *Nature*, **241**, 384.
Hoyle, F., and Narlikar, J. V. 1963, *Proc. Roy. Soc.*, *A*, **273**, 4.
Hu, B. L. 1974, *Phys. Rev.*, **D9**, 3263.

Hu, B. L., Fulling, C. A., and Parker, L. 1973, *Phys. Rev.*, **D8**, 2377.
Hubble, E. 1927, *Proc. Nat. Acad. Sci.*, **15**, 168.
Huchtmeier, W. K. 1975, *Astr. Ap.*, **45**, 259.
Hunter, C. 1962, *Ap. J.*, **136**, 594.
Iben, I., and Rood, R. T. 1970, *Ap. J.*, **159**, 605.
Illarionov, A. F., and Sunyaev, R. A. 1974*a*, *Astr. Zh.*, **51**, 698.
———. 1974*b*, *Astr. Zh.*, **51**, 1162.
Isaacson, R. A. 1968*a*, *Phys. Rev.*, **166**, 1263.
———. 1968*b*, *Phys. Rev.*, **166**, 1272.
Jacobs, K. C. 1969, preprint, Cambridge.
Jauncey, D. L. 1970, preprint.
Jeans, J. H. 1902, *Phil. Trans.*, **129**, 44.
———. 1929, *Astronomy and Cosmology* (Cambridge: Cambridge University Press); republished in 1969 (London and New York: Cambridge University Press).
Jones, B. J. T., and Steigman, G. 1978, *M.N.R.A.S.*, **183**, 585.
Jordan, P. 1948, *Astr. Nachr.*, **276**, 1955.
———. 1955, *Schwerkraft und Weltall* (Braunschweig: Vieweg und Sohn).
Kacharov, G. E. 1965, *Izv. Akad. Nauk SSSR*, **5**, 795.
Kantowski, R. 1969, *Ap. J.*, **155**, 89.
Kaplan, S. A., and Pikel'ner, S. B. 1963, *Mezhzvezdnaya Sreda* (Interstellar Medium) (Moscow: Fizmatgiz).
Karachentsev, I. D. 1968, *Rept. Byurakan Obs.*, **39**, 76.
Kardashev, N. S. 1967, *Astr. Tsirk.*, No. 430.
Kasner, E. 1921, *Amer. J. Math.*, **43**, 217.
Kawabata, K., Fujimoto, M., Sofue, Y., and Fukui, M. 1969, *Pub. Astr. Soc. Japan*, **21**, 293.
Kellermann, K. I. 1972, *Astr. J.*, **77**, 531.
Kenderdine, S., Ryle, M., and Pooley, G. G. 1966, *M.N.R.A.S.*, **134**, 189.
Khalatnikov, I. M. 1965, *Zh. Eksp. Teoret. Fiz.*, **48**, 261.
———. 1967, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **5**, 195.
Khalatnikov, I. M., and Lifshitz, E. M. 1970, *Phys. Rev. Letters*, **24**, 76.
Kholopov, A. S. 1962, ed., *Stroenie Zvezdnykh Sistem* ("Structure of Stellar Systems") (Leningrad: I. L.).
Kipper, A. 1950, *Astr. Zh.*, **27**, 321.
Kirzhnits, D. A. 1973, *Priroda*, No. 1, p. 38.
Kirzhnits, D. A., and Fainberg, V. Ya. 1973, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **18**, 596.
Kirzhnits, D. A., and Linde, A. D. 1974, *Zh. Eksp. Teoret. Fiz.*, **67**, 1263.
Kislyakov, A. G., Chernyshev, V. I., Lebskiy, Yu. V., Mal'tsev, V. A., and Serov, N. V. 1971, *Astr. Zh.*, **48**, 39.
Klein, O. 1961, *Werner Heisenberg und die Physik unserer Zeit*, S. 58, (Braunschweig: Keweg).
Klimov, Yu. G. 1963, *Astr. Zh.*, **40**, 874.

Kobzarev, I. Yu., and Okun', L. B. 1962, *Zh. Eksp. Teoret. Fiz.*, **43**, 1904.
Kobzarev, I. Yu., and Peshkov, P. I. 1974, *Zh. Eksp. Teoret. Fiz.*, **67**, 428.
Kolmogorov, A. N. 1941, *Doklady Akad. Nauk SSSR*, **30**, 299.
Komberg, B. V., and Novikov, I. D. 1975, *Pis'ma Astr. Zh.*, **1**, No. 3, p. 3.
Komberg, B. V., and Ruzmaykin, A. A. 1972, preprint, Institute of Applied Mathematics, No. 8.
Kompaneets, A. S. 1956, *Zh. Eksp. Teoret. Fiz.*, **31**, 876.
Kompaneets, A. S., and Chernov, A. S. 1964, *Zh. Eksp. Teoret. Fiz.*, **47**, 1939.
Konstantinov, B. P., Bredov, M. M., Belyavsky, A. I., and Sokolov, I. A. 1966, *Kosmicheskie Issledovaniya*, **4**, 66.
Kormendy, J., and Bahcall, J. N. 1973, *Astr. J.*, **79**, 671.
Kunkel, W. E., and Demers, S. 1976, *R. Greenwich Obs. Bull.* No. 182, p. 241.
Kurskov, A. A., and Ozernoy, L. M. 1974*a*, *Astr. Zh.*, **51**, 270.
———. 1974*b*, *Astr. Zh.*, **51**, 508.
———. 1974*c*, *Astr. Zh.*, **51**, 1177.
Kurt, V. G., and Sunyaev, R. A. 1967*a*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **5**, 299.
———. 1967*b*, *Kosmicheskie Issledovaniya*, **5**, 573.
———. 1967*c*, *Astr. Zh.*, **44**, 6.
———. 1970, *Proc. IAU Symposium 36*, Holland.
Kuz'min, V. A. 1970, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **13**, 335.
Lake, G., and Partridge, R. B. 1977, *Nature*, **270**, 502.
Landau, L. D., and Lifshitz, E. M. 1953, *Mekhanika Sploshnikh Sred* (Fluid Mechanics) (Moscow: Gostekhizdat).
———. 1957, *Zh. Eksp. Teoret. Fiz.*, **32**, 618.
———. 1964, *Statisticheskaya Fizika* (Statistical Physics) (Moscow: Fizmatgiz).
———. 1962, *Teoriya Polya* (Classical Theory of Fields) (Moscow: Nauka).
Landau, L. D., and Pomeranchuk, I. Ya. 1955, *Doklady Akad. Nauk SSSR*, **102**, 489.
Layzer, D. 1968, *Ap. Letters*, **1**, 99.
Layzer, D., and Hively, R. 1973, *Ap. J.*, **179**, 361.
Lemaitre, G. 1927, *Ann. Soc. Sci. Bruxelles*, **A47**, 49.
———. 1933, *Ann. Soc. Sci. Bruxelles*, **A53**, 51.
Levich, E. V., and Sunyaev, R. A. 1971, *Astr. Zh.*, **48**, 461.
Lifshitz, E. M. 1946, *Zh. Eksp. Teoret. Fiz.*, **16**, 587.
Lifshitz, E. M., and Khalatnikov, I. M. 1960, *Zh. Eksp. Teoret. Fiz.*, **39**, 149.
———. 1963*a*, *Uspekhi Fiz. Nauk*, **80**, 391.
———. 1963*b*, *Adv. in Physics*, **12**, 185.
Lifshitz, E. M., Lifshitz, I. M., and Khalatnikov, I. M. 1970, *Zh. Eksp. Teoret. Fiz.*, **59**, 322.
Lillie, C. F. 1969, *Bull. Amer. Astr. Soc.*, **1**, 132.
Linde, A. D. 1974, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **19**, 320.
Lindquist, R. W. 1966, *Ann. Phys.*, **37**, 487.

Ljubimov, V. A., Novikov, E. G., Nosik, V. S., Tretyakov, E. F., and Kosic, V. S. 1980, *Yad. Fiz.*, Vol. **32**, No. 7.

Longair, M. S. 1966, *M.N.R.A.S.*, **133**, 421.

———. 1974, ed., *IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data* (Dordrecht: Reidel).

Longair, M. S., and Rees, M. J. 1972, *Comm. Ap. Space Phys.*, **3**, 79.

Longair, M. S., and Sunyaev, R. A. 1969*a*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **10**, 56.

———. 1969*b*, *Nature*, **223**, 719.

———. 1969*c*, *Ap. Letters*, **4**, 65.

———. 1971, *Uspekhi Fiz. Nauk*, **105**, 41.

Low, F. J., and Tucker, W. H. 1968, *Phys. Rev. Letters*, **21**, 1538.

Lukash, V. N. 1973, *Zh. Eksp. Teoret. Fiz.*, **67**, 1594.

———. 1974*a*, *Astr. Zh.*, **51**, 281.

———. 1974*b*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **19**, 499.

———. 1974*c*, *Zh. Eksp. Teoret. Fiz.*, **67**, 1594.

Lukash, V. N., Novikov, I. D., and Starobinsky, A. A. 1975, *Zh. Eksp. Teoret. Fiz.*, **69**, 1484.

Lukash, V. N., and Starobinsky, A. A. 1974, *Zh. Eksp. Teoret. Fiz.*, **66**, 1515.

Lynden-Bell, D. 1962, *Proc. Cambridge Phys. Soc.*, **58**, 709.

———. 1964, *Ap. J.*, **139**, 1995.

———. 1976, *M.N.R.A.S.*, **174**, 695.

MacCallum, M. A. N. 1971, *Comm. Math. Phys.*, **20**, 57.

Magalinsky, V. B., and Sil'vestrov, V. V. 1972, *Tezisy Dokladov III Sovetskoi Gravitatsionnoi Konferentsii* (Proceedings of the Third Soviet Gravitation Conference) (Yerevan).

Mamaev, S. G., Masstepanenko, V. M., and Starobinsky, A. A. 1976, in *Proceedings of All-Soviet Conference on Gravitation* (Minsk, USSR); also *Zh. Eksp. Teoret. Fiz.*, **70**, 1577.

Markov, M. A. 1966, *Zh. Eksp. Teoret. Fiz.*, **51**, 878.

———. 1973, *Doklad Na 6-i Gravitatsionnoi Konferentsii* (Proceedings of the 6th Gravitation Conference) (Yerevan).

Marochnik, L. S., and Suchkov, A. A. 1974, *Uspekhi Fiz. Nauk*, **112**, 275.

Marx, G. 1973, in Proc. *IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).

———. 1977, ITP-376 preprint, Budapest.

Marx, G., and Szalay, A. S. 1972, *Proc. "Neutrino 72"* (Budapest: Technoinform), Vol. **1**, p. 123.

Mathewson, D. S., Cleary, M. N., and Murray, J. D. 1975, *Ap. J.* (*Letters*), **195**, L97.

Mattig, W. 1958, *Astr. Nachr.*, **284**, 109.

———. 1959, *Astr. Nachr.*, **285**, 1.

Matzner, R. 1968, *Ap. J.*, **154**, 1123.

Matzner, R., and Misner, C. W. 1972, *Ap. J.*, **171**, 415.

Matzner, R., Ryan, M., and Bion, R. 1973, preprint.
Matzner, R., Shepley, L. C., and Warren, J. B. 1970, *Ann. Phys.*, **57**, 401.
McCone, A. J. 1971, *Nucl. Sci. Abstr.*, **25**, No. 15.
McCrea, W. H. 1951, *Proc. Roy. Soc.*, *A*, **206**, 562.
McCrea, W. H., and Milne, E. 1934, *Quart. J. Math.*, **5**, 73.
McKellar, A. 1941, *Pub. Dominion Ap. Obs.*, **7**, No. 15.
McVittie, G. C. 1956, *General Relativity and Cosmology* (London); 2d ed. 1965 (Urbana: University of Illinois).
———. 1959*a*, in *Handbuch der Physik*, ed. S. Flügge (Berlin: Springer), p. 53.
———. 1959*b*, in *Les Theories relativistes de la gravitation*, Proceedings of 1959 Royaumont conference (Paris: CNRS, 1962).
———. 1962*a*, *Fact and Theory in Cosmology* (New York: Macmillan).
———. 1962*b*, *Phys. Rev.*, **128**, 2871.
Mikhailovskiy, A. B., and Friedman, A. M. 1971, *Zh. Eksp. Teoret. Fiz.*, **61**, 457.
———. 1973, *Astr. Zh.*, **50**, 88.
Mikhailovskiy, A. B., Friedman, A. M., and Epel'baum, Ya. G. 1970, *Zh. Eksp. Teoret. Fiz.*, **59**, 1608.
Millea, M. F., McColl, M., Pederson, R. J., and Vernon, F. L. 1971, *Phys. Rev. Letters*, **26**, 919.
Milne, E. A. 1934, *Quart. J. Math. Oxford*, **5**, 64.
———. 1935, *Relativity, Gravitation and World-Structure* (London and New York: Oxford University Press).
———. 1948, *Kinematic Relativity* (London and New York: Oxford University Press).
Mishustin, I. N., and Ruzmaykin, A. A. 1971, *Zh. Eksp. Teoret. Fiz.*, **61**, 441.
Misner, C. W. 1967, *Phys. Rev. Letters*, **19**, 533.
———. 1968, *Ap. J.*, **158**, 431.
———. 1969*a*, *Phys. Rev. Letters*, **22**, 1071.
———. 1969*b*, *Phys. Rev.*, **186**, 1319, 1328.
Misner, C. W., Thorne, K. S., and Wheeler, J. A. 1973, *Gravitation* (San Francisco: W. H. Freeman and Company).
Monin, A. S., and Yaglom, A. M. 1965, *Statisticheskaya Gidromekhanika, Ch. 1* (Statistical Hydromechanics, Part 1) (Moscow: Nauka).
———. 1967, *Statisticheskaya Gidromekhanika, Ch. 2* (Statistical Hydrodynamics, Part 2) (Moscow: Nauka).
Morgenstern, R. E. 1973, *Phys. Rev.*, **D7**, 1570.
Muehlner, D. J., and Weiss, R. 1970, *Phys. Rev. Letters*, **24**, 742.
———. 1973, *Phys. Rev. Letters*, **30**, 757.
Nariai, H. 1969*a*, *Progr. Theor. Phys.*, **42**, 544.
———. 1969*b*, *Progr. Theor. Phys.*, **42**, 742.
———. 1974, *Progr. Theor. Phys.*, **51**, 613.

Narlikar, J. V. 1963, *M.N.R.A.S.*, **126**, 203.
Narozhny, N. B., and Nikishov, A. I. 1970, *Yadern Fiz.*, **11**, 1072.
———. 1973, *Zh. Eksp. Teoret. Fiz.*, **65**, 862.
Ne'eman, Y. 1965, *Ap. J.*, **141**, 1303.
Neyman, J., and Scott, E. L. 1962, *Ap. J.*, **67**, 582.
Nikishov, A. I. 1969, *Zh. Eksp. Teoret. Fiz.*, **57**, 1210.
Nordsieck, K. H. 1973, *Ap. J.*, **184**, 719.
Novikov, I. D. 1961, *Astr. Zh.*, **38**, 564.
———. 1962, *Vestnik, Moscow State University*, ser. 3, No. 6, 61.
———. 1963, *Astr. Zh.*, **40**, 772.
———. 1964*a*, *Publ. Shternberg Astron. Inst.*, No. 132, 3, 43.
———. 1964*b*, *Zh. Eksp. Teoret. Fiz.*, **46**, 686.
———. 1964*c*, *Astr. Zh.*, **41**, 1075.
———. 1968, *Astr. Zh.*, **45**, 538.
———. 1970, *Astr. Zh.*, **47**, 1203.
———. 1975, *Astr. Zh.*, **52**, 1000.
———. 1976, preprint 268, Space Research Institute (Moscow).
Novikov, I. D., and Zel'dovich, Ya. B. 1967, *Ann. Rev. Astr. Ap.* **5**, 627.
———. 1973, *Ann. Rev. Astr. Ap.*, **11**, 387.
Novikov, S. P. 1972, *Zh. Eksp. Teoret. Fiz.*, **62**, 1977.
Novokreshchenova, S. I., and Rudnitsky, G. M. 1973, *Astr. Zh.*, **50**, 877.
Nutku, Y. 1969, *Ap. J.*, **158**, 991.
Oke, J. B. 1967, *Ap. J.*, **147**, 901.
Okun', L. B. 1966, *Uspekhi Fiz. Nauk*, **89**, 603.
Okun', and Zel'dovich, Ya. B. 1976, *Comments Nucl. Particle Phys.*, **6**, 69.
Omnès, R. 1969, *Phys. Rev. Letters*, **23**, 38.
———. 1971*a*, *Astr. Ap.*, **10**, 228.
———. 1971*b*, *Astr. Ap.*, **11**, 450.
———. 1971*c*, *Astr. Ap.*, **15**, 275.
Oort, J. H. 1958, in *La structure et l'evolution de l'univers*, 11 Conseil de Physique Solvay (Brussels: Stoops).
———. 1970, *Astr. Ap.*, **7**, 384.
Ostriker, J. P., Peebles, P. J. E., and Yahil, A. 1974, *Ap. J.* (*Letters*), **193**, L1.
Ozernoy, L. M. 1971, *Astr. Zh.*, **48**, 1160.
———. 1974, *Astr. Zh.*, **51**, 1108.
Ozernoy, L. M., and Chernin, A. D. 1967, *Astr. Zh.*, **44**, 1131.
———. 1968, *Astr. Zh.*, **45**, 1137.
Ozernoy, L. M., and Chernomordik, V. V. 1976, *A. J.* (*USSR*), **53**, 459.
Ozernoy, L. M., and Chibisov, G. V. 1970, *Astr. Zh.*, **47**, 769.
Ozernoy, L. M., Rozental', I. L., and Prilutsky, O. F. 1973, *Astrofizika Visokikh Energii* (High Energy Astrophysics) (Moscow: Atomizdat).
Paal, G. 1971, *Acta Phys. Acad. Sci. Hungaricae*, **30**, 51.
Paradoksov, P. 1966, *Uspekhi Fiz. Nauk*, **89**, 707.
Parijsky, Yu. N. 1968, *Astr. Zh.*, **45**, 279.

. 1972, *Astr. Zh.*, **49**, 1322.
. 1973, *Astr. Zh.*, **50**, 453.
Parker, E. N. 1955, *Ap. J.*, **122**, 293.
———. 1971, *Ap. J.*, **163**, 255.
Parker, L. 1968, *Phys. Rev. Letters*, **21**, 562.
———. 1969, *Phys. Rev.*, **183**, 1057.
———. 1971, *Phys. Rev.*, **D3**, 346.
———. 1972, *Phys. Rev. Letters*, **28**, 705.
———. 1973, *Phys. Rev.*, **D7**, 976.
Parker, L., and Fulling, S. A. 1974, *Phys. Rev.*, **D9**, 341.
Partridge, R. B. 1973, in *Proc. IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
Partridge, R. B., and Peebles, P. J. E. 1967*a*, *Ap. J.*, **147**, 868.
———. 1967*b*, *Ap. J.*, **148**, 377.
Pati, J. S., and Salam, A. 1973, *Phys. Rev. Letters*, **31**, 661.
Pauli, W. 1958, *Theory of Relativity* (New York: Pergamon).
Peach, V. J. 1972, *Proc. IAU Symposium 44, Galaxies and Quasi-stellar Objects*, ed. D. S. Evans (Dordrecht: Reidel).
Pecker, J. C., Roberts, A. P., and Vigier, J. P. 1972, *Nature*, **237**, 227.
Peebles, P. J. E. 1966*a*, *Phys. Rev. Letters*, **16**, 410.
———. 1966*b*, *Ap. J.*, **146**, 542.
———. 1967*a*, *Ap. J.*, **147**, 859.
———. 1967*b*, in *Proc. 4th Texas Conference on Relativistic Astrophysics* (New York).
———. 1968, *Ap. J.*, **153**, 1.
———. 1969*a*, *Ap. J.*, **155**, 393.
———. 1969*b*, *Ap. J.*, **157**, 1075.
———. 1970, *Phys. Rev.*, **D1**, 391.
———. 1971*a*, *Physical Cosmology* (Princeton: Princeton University Press).
———. 1971*b*, *Ap. Space Sci.*, **11**, 443.
———. 1973*a*, *Ap. J.*, **180**, 1.
———. 1973*b*, *Ap. J.*, **185**, 413.
———. 1973*c*, *Publ. Astr. Soc. Japan*, **25**, 291.
———. 1974*a*, *Ap. J.* (*Letters*), **189**, L51.
———. 1974*b*, *Astr. Ap.*, **32**, 391.
Peebles, P. J. E., and Hauser, M. G. 1973, *Ap. J. Suppl.*, **252**.
Peebles, P. J. E., and Yu, J. T. 1970, *Ap. J.*, **162**, 815.
Peierls, S. E., Singwi, K. S., and Wroe, D. 1952, *Phys. Rev.*, **87**, 46.
Peimbert, M. 1968, *Bol. Obs. Tonantzintla y Tacubaya*, **30**, 233.
Pelyushenko, S. A., and Stankevich, K. S. 1969, *Astr. Zh.*, **46**, 283.
Penrose, R. 1964, in *Relativity, Groups and Topology*, ed. C. M. De Witt and B. S. De Witt, (New York: Gordon & Breach), p. 565.
———. 1965, *Phys. Rev. Letters*, **14**, 57.

———. 1966, in *Perspectives in Geometry and Relativity: Essays in Honor of Václav Hlavatý* (Bloomington: Indiana University Press).
———. 1968, *Structure of Spacetime* (New York and Amsterdam: W. A. Benjamin).
———. 1969, *Rev. Nuovo Cimento*, **1**, 252.
Penzias, A. A. 1968, *Trans. Microwave Theory and Techniques IEEE*, MTT-16, **9**, 608.
Penzias, A. A., and Wilson, R. W. 1965, *Ap. J.*, **142**, 419.
———. 1967, *Astr. J.*, **72**, 315.
Peres, A. 1960, *Progr. Theor. Phys.*, **24**, 149.
Petrosian, V. 1974, in *IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
Petrosian, V., Salpeter, E., and Szekeres, P. 1967, *Ap. J.*, **147**, 1222.
Petrov, A. Z. 1966, *Novie Metody v Obshchei Teorii Otnositel'nosti* (New Methods in the General Theory of Relativity) (Moscow: Nauka).
Peyraud, J. 1968, *J. de Phys.*, **29**, 88, 306, 872.
Pikel'ner, S. B. 1967, *Astr. Zh.*, **44**, 915.
———. 1974, *Comm. Ap. Space Phys.*, **6**, 15.
Pikel'ner, S. B., and Vainshtein, L. A. 1966, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **4**, 307.
Pokrovskii, G. I., and Korsunskii, V. K. 1931, *Naturwiss.*, **19**, 421.
Pontecorvo, B. M., and Smorodinskii, Ya. A. 1961, *Zh. Eksp. Teoret. Fiz.*, **41**, 239.
Press, W. H., and Schechter, P. 1974, *Ap. J.*, **187**, 425.
Press, W. H., and Thorne, K. S. 1972, *Ann. Rev. Astr. Ap.*, **10**, 335.
Puzanov, V. I., Salomonovich, A. E., and Stankevich, K. S. 1967, *Astr. Zh.*, **44**, 1129.
Reber, G. R. 1968, *J. Franklin Institute*, **285**, 1.
Rees, M. J. 1971, *M.N.R.A.S.*, **154**, 187.
———. 1972, *Phys. Rev. Letters*, **28**, 1669.
Reeves, H. 1973, preprint SEP (SES) 73-77R HR.
———. 1974, *Ann. Rev. Astr. Ap.*, **12**, 437.
Reeves, H., Audouze, J., Fowler, W. A., and Schramm, D. N. 1973, *Ap. J.*, **179**, 909.
Refsdal, S., and Stabell, R. 1966, *M.N.R.A.S.*, **132**, 379.
Reines, E., Sobel, H. W., and Pasierb, E. 1980, University of California preprint.
Reinhardt, M. 1972, *M.N.R.A.S.*, **156**, 151.
Reinhardt, M., and Roberts, M. S. 1972, *Ap. Letters*, **12**, 201.
Reinhardt, M., and Thiel, M. 1970, *Ap. Letters*, **7**, 101.
Richards, P. L. 1975, *Bull. Am. Phys. Soc.*, **20**, 4, 540.
Roach, F. E., and Smith, L. L. 1968, *Geophys. J.*, **15**, 227.
Roberts, M. S., and Whitehurst, R. N. 1975, *Ap. J.*, **201**, 327.
Robertson, H. P. 1955, *Publ. Astr. Soc. Pacific*, **67**, No. 395, 82.

Robson, E. I., Vicker, D. G., Huizinga, J. S., Beckmann, I. E., and Clegg, P. E. 1974, *Nature*, **251**, 591.
Rogerson, Y. B., and York, D. C. 1973, *Ap. J.* (*Letters*), **186**, L95.
Roll, P. G., and Wilkinson, D. T. 1966, *Phys. Rev. Letters*, **16**, 405.
———. 1967, *Ann. Phys.*, **44**, 289.
Rood, H., Page, T. L., Kintner, E. C., and King, I. K. 1972, *Ap. J.*, **175**, 627.
Rosen, G. 1964, *Phys. Rev.*, **136**, 2798.
Rozhansky, V. 1940*a*, *Ap. J.*, **91**, 257.
———. 1940*b*, *Phys. Rev.*, **58**, 1010.
Ruban, V. A. 1971, preprint, LFTI, No. 355.
Ruban, V. A., and Finkel'shteyn, A. M. 1973, Preprint, L.I.Ya.F., No. 59.
Ruzmaykina, T. V. 1972, *Astr. Zh.*, **49**, 1229.
Ruzmaykina, T. V., and Ruzmaykin, A. A. 1969, *Zh. Eksp. Teoret. Fiz.*, **57**, 680.
Ryan, M. 1972, *Hamiltonian Cosmology* (Berlin, Heidelberg, New York: Springer-Verlag).
Sachs, R. K., and Wolfe, A. M. 1967, *Ap. J.*, **147**, 73.
Sakharov, A. D. 1965, *Zh. Eksp. Teoret. Fiz.*, **49**, 345.
———. 1966, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **3**, 439.
———. 1967*a*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **5**, 32.
———. 1967*b*, *Doklady Akad. Nauk SSSR*, **177**, 70.
Salpeter, E. E. 1971, *Nature Phys. Sci.*, **233**, 5.
Sandage, A. R. 1961, *Ap. J.*, **133**, 355.
———. 1962, in *Proc. IAU Symposium 15, Problems of Extragalactic Research*, ed. G. C. McVittie (New York: Macmillan), p. 359.
———. 1968, *Observatory*, **33**, 91.
———. 1970, *Physics Today*, **23**, 44.
———. 1972*a*, *Ap. J.*, **173**, 485.
———. 1972*b*, *Ap. J.*, **178**, 1.
———. 1972*c*, *Ap. J.*, **178**, 25.
———. 1973*a*, *Ap. J.*, **183**, 711.
———. 1973*b*, *Ap. J.*, **183**, 731.
Sandage, A. R., Sandage, M., and Kristian, J. 1975, eds., *Stars and Stellar Systems*, Vol. **9**, *Galaxies and the Universe* (Chicago: University of Chicago Press).
Sandage, A. R., and Tammann, G. A. 1974*a*, *Ap. J.*, **190**, 525.
———. 1974*b*, *Ap. J.*, **191**, 603.
———. 1974*c*, *Ap. J.*, **194**, 223.
———. 1974*d*, *Ap. J.*, **194**, 559.
———. 1975*a*, *Ap. J.*, **196**, 313.
———. 1975*b*, *Ap. J.*, **197**, 265.
Sato, H. 1971, *Progr. Theor. Phys.*, **45**, 370.
Sato, H., Matsuda, T., and Takeda, H. 1970, *Progr. Theor. Phys.*, **43**, 1115.

Savedoff, M. P., and Vila, S. 1962, *Ap. J.*, **136**, 609.
Schlüter, A., and Biermann, L. 1950, *Zs. Naturforsch.*, **5a**, 237.
Schramm, D. N. 1974, *Ann. Rev. Astr. Ap.*, **12**, 383.
Schücking, E. 1963, in *Gravitation. An Introduction to Current Research*, ed. L. Witten (New York: Wiley), p. 460.
Schücking, E., and Heckmann, O. 1958, in *La Structure et l'Evolution de l'Univers*, 11 Conseil de Physique Solvay (Brussels: Stoops).
Schwartz, D. 1970, *Ap. J.*, **162**, 439.
Sciama, D. W. 1971, *Modern Cosmology* (Cambridge: Cambridge University Press).
Searle, L., and Sargent, W. L. W. 1972, *Comments Ap. and Space Phys.*, **4**, 59.
Seldner, M., Siebers, B., Groth, E. J., and Peebles, P. J. E. 1977, *Astr. J.*, **82**, 249.
Setti, G., and Woltjer, L. 1973, *Ap. J.* (*Letters*), **181**, L61.
Shandarin, S. F. 1974, *Astr. Zh.*, **51**, 667.
Shapiro, I. 1965, *Icarus*, **4**, 549.
———. 1972, *J. General Relativity*, **1**, 135.
Shapiro, I., Petengill, G., Ash, M., Ingolls, R., Campbell, D., and Dicke, R. 1972, *Phys. Rev. Letters*, **28**, 1597.
Shapiro, I., Smith, W., Ash, M., Ingolls, R., and Petengill, G. 1971, *Phys. Rev. Letters*, **26**, 27.
Shapiro, I. S. 1973, *Uspekhi Fiz. Nauk*, **109**, 432.
Shikin, I. S. 1970, *Zh. Eksp. Teoret. Fiz.*, **59**, 182.
Shivanandan, K., Houck, J. R., and Harwit, M. 1968, *Phys. Rev. Letters*, **21**, 1460.
Shklovsky, I. S. 1965, *Astr. Zh.*, **42**, 893.
———. 1967, *Astr. Tsirk.*, No. 429.
Shrum, E. B., and Ziock, K. O. H. 1971, *Phys. Letters*, **37B**, 115.
Shvartsman, V. F. 1969, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **9**, 315.
Sibgatulin, N. R. 1974, *Zh. Eksp. Teoret. Fiz.*, **66**, 1187.
Silk, J. 1968, *Ap. J.*, **151**, 459.
Sil'vestrov, V. V. 1974, *Astr. Zh.*, **51**, 293.
Sokolov, D. D. 1970, *Doklady Akad. Nauk SSSR*, **195**, 1307.
Sokolov, D. D., and Shvartsman, V. F. 1974, *Zh. Eksp. Teoret. Fiz.*, **66**, 412.
Sokolov, D. D., and Starobinsky, A. A. 1975, *Astr. Zh.*, **52**, 1041.
Spinrad, H., and Taylor, B. Z. 1971, *Ap. J.*, **22**, 445.
Spitzer, L. 1962, *Physics of Fully Ionized Gases* (New York: Wiley).
Spitzer, L., and Greenstein, J. L. 1951, *Ap. J.*, **114**, 407.
Spitzer, L. J., and Tomasko, M. G. 1968, *Ap. J.*, **152**, 971.
Sretenskiy, L. N. 1946, *Teoriya N'yutonovskogo Potentsiala* (Theory of the Newtonian Potential) (Moscow: Gostekhizdat).
Stabell, R., and Refsdal, S. 1966, *M.N.R.A.S.*, **132**, 379.

Stanyukovich, K. P. 1965, *Gravitatsionnoe Pole i Elementarniye Chastitsy* (The Gravitational Field and Elementary Particles) (Moscow: Atomizdat).
———. 1966*a*, *Uspekhi Fiz. Nauk*, **89**, 731.
———. 1966*b*, *Doklady Akad. Nauk SSSR*, **168**, 781.
———. 1971, *Neustanovivshiesya Dvizheniya Sploshnoi Sredi* (Nonstationary Motion of Continuous Media) (Moscow: Nauka).
Stanyukovich, K. P., and Sharshikeyev, O. Sh. 1970, *Astrofizika*, **6**, 571.
———. 1973, *P.M.M.*, **37**, 739.
Starobinsky, A. A. 1973, *Zh. Eksp. Teoret. Fiz.*, **64**, 48.
———. 1975, in *Proceedings of First Marcel Grossman Meeting on General Relativity*, ed. R. Ruffini (Amsterdam: North Holland).
———. 1976, preprint.
Stecker, F. W. 1969, *Nature*, **224**, 870.
Steenbeck, M. 1963, *Monatsbereichte Dtsch. Akad. Wiss. Berlin*, **5**, 625.
Steenbeck, M., Krause, F., and Rädler, K. H. 1966, *Zs. Naturforsch.*, **21a**, 369.
Steigman, G. 1971, in *General Relativity and Cosmology*, ed. R. K. Sachs (New York: Academic Press).
———. 1973, in *Proc. IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
———. 1976, *Ann. Rev. Astr. Ap.*, **14**, 339.
Stokes, R. A., Partridge, R. B., and Wilkinson, D. T. 1967, *Phys. Rev. Letters*, **19**, 199.
Sunyaev, R. A. 1968, *Doklady Akad. Nauk SSSR*, **179**, 45.
———. 1969*a*, *Astr. Zh.*, **46**, 929.
———. 1969*b*, *Ap. Letters*, **3**, 33.
———. 1971, *Astr. Zh.*, **48**, 244.
Sunyaev, R. A., and Zel'dovich, Ya. B. 1969, *Nature*, **223**, 721.
———. 1970*a*, *Ap. Space Sci.*, **7**, 3.
———. 1970*b*, *Ap. Space Sci.*, **7**, 20.
———. 1970*c*, *Comments Ap. Space Phys.*, **2**, 66.
———. 1972*a*, *Comments Ap. Space Phys.*, **4**, 173.
———. 1972*b*, *Astr. Ap.*, **20**, 189.
———. 1980, *Ann. Rev. Astr. Ap.*, **18**, 489.
Süveges, M. 1966, *Acta Phys. Acad. Sci. Hungaricae*, **20**, 273.
Suzuki, S. 1931, *Proc. Math. Soc. Japan*, **13**, 277.
Szalay, A. S., and Marx, G. 1976, *Astr. Ap.*, **49**, 437.
Tagirov, E. A., and Chernikov, N. A. 1968, preprint, O.I.Ya.I. P2-3777.
Talbot, R. J., and Arnett, W. D. 1973*a*, *Ap. J.*, **186**, 51.
———. 1973*b*, *Ap. J.*, **186**, 69.
Tammann, G. A. 1973*a*, *Proc. IAU Symp. No. 63*, Poland.
———. 1973*b*, preprint, Astronom. Inst. der Univ. Basel und Hamburg. Sternwartl.

Tarter, J., and Silk, J. 1974, *Quart. J.R.A.S.*, **15**, 122.
Taub, A. 1951, *Ann. Math.*, **53**, 472.
Teukolsky, S. A., and Press, W. H. 1974, *Ap. J.*, **193**, 443.
Thaddeus, P. 1972, *Ann. Rev. Astr. Ap.*, **10**, 305.
Thaddeus, P., and Clauser, J. F. 1966, *Phys. Rev. Letters*, **16**, 819.
t'Hooft, G. 1976, *Phys. Rev. Letters*, **37**, 8.
t'Hooft, G., and Veltman, M. 1974, *Ann. Inst. Henri Poincaré*, **20**, 69.
Thorne, K. S. 1967, *Ap. J.*, **148**, 51.
Thuan, T. X., and Ostriker, J. P. 1974, *Ap. J.* (*Letters*), **191**, L105.
Tifft, W. G., and Gregory, S. A. 1976, *Ap. J.*, **205**, 696.
Tinsley, B. M. 1968, *Ap. J.*, **151**, 547.
———. 1972, *Ap. J.* (*Letters*), **173**, L93.
———. 1973, *Ap. J.*, **186**, 35.
Tinsley, B. M., and Spinrad, H. 1971, *Ap. Space Sci.*, **12**, 118.
Tiunov, E. A., and Chernin, A. D. 1971, *Astrofizika*, **7**, 161.
Tolman, R. C. 1922, *J. Am. Chem. Soc.*, **49**, 1902.
———. 1930, *Phys. Rev.*, **35**, 875.
———. 1934, *Relativity, Thermodynamics, and Cosmology* (Oxford: Clarendon Press).
Tomita, K. 1973, *Progr. Theor. Phys.*, **50**, 1283.
Tomita, K., Nariai, M., Sato, H., Matsuda, T., and Takeda, H. 1971, *Progr. Theor. Phys.*, **43**, 1511.
Trautman, A. 1973, *Nature*, **242**, 7.
Tremaine, S., and Gunn, J. E. 1979, *Phys. Rev. Letters*, **42**, 407.
Trimble, V. L. 1975, *Rev. Mod. Phys.*, **47**, 877.
Truran, J. W., and Cameron, A. G. W. 1971, *Ap. Space Sci.*, **14**, 179.
Turner, E. L. 1976, *Ap. J.*, **208**, 20.
Turner, E. L., and Gott, J. R. 1976, *Ap. J.*, **209**, 6.
Tyson, J. A. 1973, *Proc. IAU Symposium 63, Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
Vainshtein, S. I. 1971, *Zh. Eksp. Teoret. Fiz.*, **61**, 612.
Vainshtein, S. I., and Ruzmaykin, A. A. 1971, *Astr. Zh.* **48**, 918.
———. 1972, *Astr. Zh.*, **49**, 448.
———. 1973, *Astr. Zh.*, **50**, 12.
Vainshtein, S. I., and Zel'dovich, Ya. B. 1972, *Uspekhi Fiz. Nauk*, **106**, 431.
Vajk, J. P. 1969, *J. Math. Phys.*, **10**, 1145.
van den Bergh, S. 1960, *Publ. Dunlap Obs.*, **2**, No. 6.
Verschuur, G. L. 1975, *Ann. Rev. Astr. Ap.*, **13**, 257.
Vette, J. I., Gruber, D., Matteson, J. L., and Peterson, L. E. 1970, *Ap. J.* (*Letters*), **60**, L161.
Vladimirov, Yu. S. 1963, *Zh. Eksp. Teoret. Fiz.*, **45**, 251.
von Laue, M. 1931, *Sitzber. Preuss. Akad. Wiss.*, 723.
von Weizsäcker, C. F. 1951, *Ap. J.*, **114**, 165.

Vorontsov Vel'yaminov, B. A. 1972, *Vnegalakticheskaya Astronomiya* (Extragalactic Astronomy) (Moscow: Nauka).
Wagoner, R. V. 1973, *Ap. J.*, **179**, 343.
Wagoner, R. V., Fowler, W. A., and Hoyle, F. 1967, *Ap. J.*, **148**, 3.
Weber, J. 1969, *Phys. Rev. Letters*, **22**, 1320.
Weinberg, S. 1962, *Phys. Rev.*, **128**, 1457.
———. 1971, *Ap. J.*, **168**, 175.
———. 1972, *Gravitation and Cosmology* (New York: Wiley).
Welch, W. J., Keachie, S., Thornton, D. D., and Wrixon, G. 1967, *Phys. Rev. Letters*, **18**, 1068.
Weymann, R. 1965, *Phys. Fluids*, **8**, 2112.
———. 1967, *Ap. J.*, **147**, 887.
Wheeler, J. A. 1958, in *La structure et l'evolution de l'univers*, 11 Conseil de Physique Solvay (Brussels: Stoops).
———. 1960, *Neutrinos, Gravitation, and Geometry* (Bologna); see also *Geometrodynamics* (New York: Academic Press).
———. 1964, in *Gravitation and Relativity*, ed. H. Y. Chiu and W. F. Hoffman (New York: W. A. Benjamin).
———. 1968, *Einstein's Vision* (Berlin: Springer-Verlag).
Wheeler, J. A., and Feynman, R. P. 1945, *Rev. Mod. Phys.*, **17**, 157.
Whitford, A. E. 1971, *Ap. J.*, **169**, 215.
Whittaker, E. T. 1955, *Proc. Roy. Soc.*, *A*, **149**, 384.
Wilkinson, D. T. 1967*a*, *Phys. Rev. Letters*, **19**, 1195.
———. 1967*b*, *Phys. Rev. Letters*, **19**, 1251.
Williamson, K. D., Blair, A. G., Catlin, L. L., Hiebert, R. D., Loyd, E. G., and Romero, H. V. 1973, *Nature PS*, **241**, 79.
Wilson, R. W., Penzias, A. A., Jefferts, K. B., and Solomon, P. M. 1973, *Ap. J.* (*Letters*), **179**, L107.
Winter, A. J. B. 1975. *M.N.R.A.S.*, **172**, 1.
Winter, K. 1979, in *Proceedings of the International Symposium on Lepton and Photon Interactions*, ed. T. B. W. Kirk and H. D. Abarbanel (Batavia, Ill.: FNAL), p. 258.
Wolf, J. A. 1967, *Spaces of Constant Curvature* (New York: McGraw-Hill).
Woody, D. D., Mather, I. C., Nishioka, N. S., and Richards, P. L. 1975, *Phys. Rev. Letters*, **34**, 1036.
Wright, M. C. H. 1973, *Ap. J.*, **179**, 453.
Yu, J. T., and Peebles, P. J. E. 1969, *Ap. J.*, **158**, 103.
Zakharov, V. E., and Sagdeev, R. Z. 1970, *Doklady Akad. Nauk SSSR*, **192**, 297.
Zel'dovich, Ya. B. 1949, *Zh. Tekh. Fiz.*, **19**, 1119.
———. 1961, *Zh. Eksp. Teoret. Fiz.*, **41**, 1609.
———. 1962*a*, *Zh. Eksp. Teoret. Fiz.*, **43**, 1037.
———. 1962*b*, *Zh. Eksp. Teoret. Fiz.*, **43**, 1561.

———. 1963, *Uspekhi Fiz. Nauk*, **80**, 357.
———. 1964*a*, *Astr. Zh.*, **41**, 19.
———. 1964*b*, *Astr. Zh.*, **41**, 873.
———. 1965*a*, *Doklady Akad. Nauk SSSR*, **163**, 1359.
———. 1965*b*, *Zh. Eksp. Teoret. Fiz.*, **48**, 986.
———. 1965*c*, *Uspekhi Fiz. Nauk*, **86**, 303.
———. 1965*d*, *Advances Astr. Ap.*, **3**, 241.
———. 1967, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **6**, 1050.
———. 1968, *Uspekhi Fiz. Nauk*, **95**, 209.
———. 1969, *Astr. Zh.*, **46**, 775.
———. 1970*a*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **11**, 30.
———. 1970*b*, *Astrofizika*, **6**, 119.
———. 1970*c*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **12**, 443.
———. 1971, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **14**, 270.
———. 1972*a*, *Zh. Eksp. Teoret. Fiz.*, **62**, 2076.
———. 1972*b*, *M.N.R.A.S.*, **160**, 1.
———. 1972*c*, in *Magic without Magic: John Archibald Wheeler*, ed. J. Klauder (San Francisco: W. H. Freeman).
———. 1972*d*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **16**, 425.
———. 1973*a*, preprint, Institute of Applied Mathematics, No. 66; cf. Longair (1974).
———. 1973*b*, *Zh. Eksp. Teoret. Fiz.*, **64**, 58.
———. 1973*c*, *Zh. Eksp. Teoret. Fiz.*, **65**, 1311.
———. 1974, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **20**, 338.
———. 1975*a*, *Pis'ma Astr. Zh.*, **1**, No. 1, 10.
———. 1975*b*, *Uspekhi Fiz. Nauk*, **115**, 169.
———. 1976, *JETP Letters*, **24**, 29.
Zel'dovich, Ya. B., Illarionov, A. F., and Sunyaev, R. A. 1972, *Zh. Eksp. Teoret. Fiz.*, **62**, 1217.
Zel'dovich, Ya. B., and Kazhdan, Ya. M. 1970, *Astrofizika*, **6**, 109.
Zel'dovich, Ya. B., Kobzarev, I. Yu., and Okun', L. B. 1974, *Zh. Eksp. Teoret. Fiz.*, **67**, 3.
Zel'dovich, Ya. B., Kurt, V. G., and Sunyaev, R. A. 1968, *Zh. Eksp. Teoret. Fiz.*, **55**, 278.
Zel'dovich, Ya. B., and Levich, E. V. 1970, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **11**, 57.
Zel'dovich, Ya. B., and Myshkis, A. D. 1972, *Elementy Prikladnoi Matematiki* (Elements of Applied Mathematics) (Moscow: Nauka).
———. 1973, *Elementy Matematicheskoi Fiziki. Sreda iz Nevzaimodeistvuyushchikh Chastits* (Elements of Mathematical Physics. Media Made from Noninteracting Particles) (Moscow: Nauka).
Zel'dovich, Ya. B., and Novikov, I. D. 1966, *Astr. Zh.*, **43**, 758.
———. 1967*a*, *Astr. Zh.*, **44**, 663.

. 1967*b*, *Relyativistskaya Astrofizika* (Relativistic Astrophysics) (Moscow: Nauka). The first half of this book is an earlier Russian version of Zel'dovich and Novikov (1971); the second half is an earlier version of the present book.

———. 1967*c*, *Pis'ma Zh. Eksp. Teoret. Fiz.*, **6**, 772.

———. 1968, in *Nonstationary Events in Galaxies:* Proc. IAU Symp. (Yerevan, USSR), pp. 220 and 280.

———. 1969, *Astr. Zh.*, **46**, 960.

———. 1970, *Astrofizika*, **6**, 379.

———. 1971, *Teoriya Tyagoteniya I Evolyutsiya Zvezd* (Moscow: Nauka). English version: *Relativistic Astrophysics*, Vol. **1**, *Stars and Relativity* (Chicago: University of Chicago Press). Cited in text as Vol. **1**.

Zel'dovich, Ya. B., Novikov, I. D., and Starobinsky, A. A. 1974, *Zh. Eksp. Teoret. Fiz.*, **66**, 1897.

Zel'dovich, Ya. B., Okun', L. B., and Pikel'ner, S. B. 1965, *Uspekhi Fiz. Nauk*, **87**, 113.

Zel'dovich, Ya. B., and Pikel'ner, S. B. 1969, *Zh. Eksp. Teoret. Fiz.*, **56**, 310.

Zel'dovich, Ya. B., and Pitayevskij, L. P. 1971, *Comm. Math. Phys.*, **23**, 185.

Zel'dovich, Ya. B., and Podurets, M. A. 1964, *Doklady Akad. Nauk SSSR*, **156**, 57.

———. 1965, *Astr. Zh.*, **42**, 963.

Zel'dovich, Ya. B., and Polnarev, A. G. 1974, *Astr. Zh.*, **51**, 30.

Zel'dovich, Ya. B., and Popov, V. S. 1971, *Uspekhi Fiz. Nauk*, **105**, 403.

Zel'dovich, Ya. B., and Rayzer, Yu. P. 1966, *Fizika Udarnikh Voln i Vysokotemperaturnikh Gidrodinamicheskikh Yavlenii* (Physics of Shock Waves and High Temperature Hydrodynamical Phenomena) (Moscow: Fizmatgiz). English translation published in two volumes by Academic Press (New York and London).

Zel'dovich, Ya. B., and Sakharov, A. D. 1966, *Yadern. Fiz.*, **4**, 395.

———. 1967, *Acta Phys. Acad. Sci. Hung.*, **22**, 153.

Zel'dovich, Ya. B., and Smorodinsky, Ya. A. 1961, *Zh. Eksp. Teoret. Fiz.*, **41**, 907.

———. 1966*a*, *Uspekhi Fiz. Nauk*, **88**, 199.

———. 1966*b*, *Uspekhi Fiz. Nauk*, **89**, 734.

Zel'dovich, Ya. B., and Starobinsky, A. A. 1971, *Zh. Eksp. Teoret. Fiz.*, **61**, 2161.

———. 1976, *Zh. Eksp. Teoret. Fiz. Pis. Red.*, **24**, 616.

Zel'dovich, Ya. B., and Sunyaev, R. A. 1969, *Ap. and Space Sci.*, **4**, 285.

———. 1970, *Ap. and Space Sci.*, **6**, 358.

———. 1974, Preprint, Institute of Applied Mathematics, No. 94.

Zel'dovich, Ya. B., Sunyaev, R. A., Doroshkevich, A. G., and Kholopov, M. Yu. 1980, preprint.

Zel'manov, A. A. 1959*a*, *Doklady Akad. Nauk SSSR*, **124**, 1030.

———. 1959*b*, *Trudy 6-vo Soveshchaniya po Voprosam Kosmogonii* (Proceedings of the 6th Conference on Cosmogonical Questions) (Moscow: Publishing House of the Academy of Sciences of the USSR).

Zweig, G. 1964, CERN-TH 412 (unpublished).

Zwicky, F. 1957, *Morphological Astronomy* (Berlin).

———. 1959, in *Handbuch der Physik*, ed. S. Flügge, **53**, (Berlin: Springer).

Zwicky, F., Herzog, E., and Wild, P. 1968, *Catalogue of Galaxies and Clusters of Galaxies* (Pasadena).

AUTHOR INDEX

Abell, G. O., 406, 408, 409, 410
Agekyan, T. A., 406
Akhiezer, A. I., 594
Alexander, J. K., 127
Alfvén, H., 121, 168, 169, 353, 569, 582
Alpher, R., 122
Ambartsumyan, V. A., 400
Anderson, J. B., ii, 675
Aristotle, 633
Arnett, W. D., 84
Arons, J., 233
Arp, H., 412
Audouze, J., 175, 184, 186

Bagrov, V. G., 598
Bahcall, J. N., 80, 412, 583
Bardeen, J. M., 353, 462
Barger, V., 413
Bechi, C., 173
Bekenstein, J. D., 607, 615
Belinsky, V. A., 479, 510, 530, 532, 541, 543, 571, 577, 578, 613, 653
Belyavsky, A. I., 352
Berestetsky, V. B., 594, 603
Berezinsky, V. S., 23, 229
Berger, B. K., 555, 593
Bertola, F., 412
Bertotti, B., 86
Bianchi, L., 532
Biermann, L., 387
Bion, R., 653
Bisnovatyi-Kogan, G. S., 127, 264, 389, 391
Blair, A. G., 132, 194, 230
Bogolyubov, N. N., 144, 655
Bogoyavlensky, O. I., 555
Böhm-Vitense, E., 84
Bol'yai, 634
Bondi, H., 103, 622, 623
Bonner, W. B., 247, 249, 250, 292
Bonometto, S. A., 270
Bortolot, V. J., 132
Boynton, P. E., 445
Bozrikov, P. V., 598
Brachmachary, R. L., 495
Braginsky, V. B., 173, 462
Brans, C., 632, 648, 649, 651
Bredov, M. M., 352
Bridle, A. N., 127
Bronshtein, M. P., 97, 98
Brown, F. A., 410
Bulgakov, M., 632
Burbidge, E. M., 410
Burbidge, G. R., 80, 101, 119, 120, 121, 308

Callan, C. G., 3
Cameron, A. G. W., 184
Carr, B. J., ii, xvi, 303, 449, 609, 610
Cavendish, H., 635
Chadwick, J., 121
Chandrasekhar, S., 628
Chapline, G. F., 303
Chernikov, N. A., 593, 603, 663
Chernin, A. D., 253, 295, 337, 355, 356, 378, 381, 394, 396
Chernomordik, V. V., 358, 403
Chernov, A. S., 491
Chibisov, G. V., 270, 274, 280, 284, 302, 305, 355, 396, 398, 399
Chincarini, G., 410
Chiu, H. Y., 153, 168, 652
Christensen, J. H., 654
Chupka, W. A., 173
Clauser, J. F., 132
Clifford, W. K., 668
Cline, D., 413
Collins, C. B., 549, 554, 555
Cowsik, R., 413
Cristodoulou, D., 607
Cronin, J. W., 654

Danin, 121
Dashevsky, V. M., 86, 88, 89
Dautcourt, G., 466
Davis, M., 410, 414
Delbruck, M., 601
Demers, S., 412
de Vaucouleurs, G., 83, 410, 412
DeWitt, B. S., 588
DeWitt, C. M., 586
DeZotti, G., 270
Dicke, R. H., 3, 124, 278, 352, 391, 392, 427, 632, 648, 649, 650, 651, 652, 653

This index was prepared by Kip S. Thorne.

Dirac, P. A. M., 583, 595, 626, 628, 629, 650, 651
Dolgov, A. D., 144, 413
Doroshkevich, A. G., xviii, 77, 95, 120, 124, 127, 178, 221, 222, 252, 270, 278, 322, 332, 334, 339, 342, 344, 347, 351, 352, 358, 365, 369, 372, 374, 377, 380, 381, 391, 392, 399, 415, 479, 491, 495, 496, 501, 506, 510, 511, 521, 523, 530, 535, 538, 541, 545, 546, 550, 552, 554, 555, 559, 562, 563, 564
Dreicer, H., 206
Duffet-Smith, P. J., 78
Dyer, C. C., 86
Dynkin, S. D., 652

Eardley, D. M., 644
Efimov, N. V., 638
Eidel'man, E. D., 378
Einasto, Ya., 405, 408, 410, 412
Einstein, A., xxxv, 16, 20, 99, 100, 102, 205, 326, 631, 632, 634, 658, 668
Eisenhart, L. P., 530
Ellis, G. F., 21, 530, 532, 538
Elsasser, W. M., 387
Elvius, A., 353
Epel'baum, Ya. G., 264
Evans, D. S., 81, 83

Faber, S., 407
Fadeev, L. D., 588
Felten, J. E., 129
Fermi, E., 123, 140, 174
Feynman, R. P., 86, 588, 610, 611, 612, 672
Field, G. B., 132, 219, 220, 224, 225, 270, 346, 353, 406, 411, 426
Finkel'shteyn, A. M., 648, 652, 653
Fishbone, L., iii, iv, xv, xviii
Fitch, V. L., 654
Fitremann, M., 388
Fock, V. A., 666
Ford, L. H., 641
Fowler, W. A., 83, 175, 178, 182, 184, 186, 188
Frautschi, S., 583
Freeman, K. C., 407
Fricke, J., 391
Friedman, A. M., 264
Friedmann, A. A., xvii, xxxv, 100, 121
Frisch, U., 388
Fujimoto, M., 385
Fukui, M., 385
Fulling, S. A., 593, 606

Gallinaro, G., 173
Gamow, G., 122, 123, 174, 355, 394, 629, 630, 631
Gauss, K. F., 634
Gell-Mann, G., 143
Geroch, R., 570, 571
Gershtein, S. S., 156, 414
Gertsenshteyn, M. E., 471
Ginzburg, V. L., 193, 207, 229, 569, 583, 586
Gitman, D. I., 598
Gold, T., 622, 623, 671
Goldsmith, D. W., 346, 409
Golenetsky, S. V., 225, 233
Gott, J. R., 342, 393, 398, 412
Gould, R. J., 221
Greenstein, J. L., 198, 653
Gregory, S. A., 410
Grib, A. A., 593
Grishchuk, L. P., 331, 510, 530, 533, 535, 536, 537, 538, 541, 545, 546, 550, 555, 562, 621, 622, 663, 665
Groth, E. J., 410
Gull, S. F., 225
Gunn, J. E., 86, 221, 370, 413, 415
Gurevich, L. E., 648, 652, 653
Guyot, M., 262, 462

Habing, H. J., 346
Hagedorn, R., 137, 144, 147, 148, 149, 160, 162, 163, 582, 583, 584, 585, 621
Harrison, B. K., 110
Harrison, E. R., 385, 432
Harwit, M. O., 132
Hauser, M. G., 317
Hawking, S. W., 21, 287, 300, 355, 526, 530, 549, 554, 555, 570, 571, 575, 576, 607, 608, 609, 610, 615, 616, 653, 663
Hayakawa, S., 229
Hayashi, C., 123, 174, 175, 177
Heckmann, O., 61, 479, 487, 526, 530, 550
Heisenberg, W., 586
Herman, R., 122
Herzog, E., 408
Hewish, A., 78
Hills, R. E., 80
Hitchcock, J. L., 132
Hively, R., 355
Holmberg, E., 406
Houck, J. R., 132
Hoyle, F., 85, 121, 175, 178, 186, 188, 389, 495, 582, 622, 624, 625, 667, 668, 671
Hu, B. L., 593
Hubble, E., xvii, 79, 83, 100, 121
Huchtmeier, W. K., 407
Hunter, C., 251

Iben, I., 83
Il'f, 583
Illarionov, A. F., 210, 213
Isaacson, R. A., 158, 452

Jackson, P., xvi
Jackson, R. E., 107
Jacobs, K. C., 495
Jauncey, D. L., 75
Jeans, J. H., 240, 241, 641, 642
Joëveer, M., 410
Jones, B. J. T., 220
Jordan, P., 650

Kaasik, A., 408, 412
Kacharov, G. E., 232
Kant, I., 355
Kantowski, R., 86
Kaplan, S. A., 196
Karachentsev, I. D., 410
Kardashev, N. S., 103, 104, 631
Kasner, E., 486
Kawabata, K., 385
Kazhdan, Ya. M., 326, 327
Kellermann, K. I., 75
Kenderdine, S., 72
Kepler, J., xviii
Khalatnikov, I. M., 285, 289, 292, 303, 304, 479, 487, 495, 501, 505, 506, 510, 515, 530, 532, 538, 541, 543, 570, 571, 572, 577, 578, 591, 613, 653
Kholopov, A. S., 415
Kipper, A., 198
Kirzhnits, D. A., 140, 583, 586, 656
Klein, O., 109, 121, 168, 230, 353, 582
Klimov, Yu. G., 88
Kobzarev, I. Yu., 162, 583, 619, 657
Kolmogorov, A. N., 357, 397, 398
Komberg, B. V., 385, 408, 412
Kompaneets, A. S., 192, 208, 212, 491
Konstantinov, B. P., 352
Kormendy, J., 412
Krainovich, V. Ya., 652
Krause, F., 388
Kristian, J., 406
Kunkel, W. E., 412
Kurskov, A. A., 396
Kurt, V. G., 128, 194, 195, 222, 223
Kuz'min, V. A., 613

Lagrange, 604
Lake, G., 225
Landau, L. D., xxxii, 20, 95, 101, 109, 113, 139, 149, 154, 158, 275, 299, 450, 451, 471, 484, 506, 507, 600, 627, 655
Laplace, P. S., 355
Latonsky, 632
Layzer, D., 119, 355
Lemaitre, G., xxxv, 121, 622, 623, 631, 637
Levich, E. V., 206, 207
Lifshitz, E. M., 20, 95, 101, 109, 113, 139, 154, 158, 285, 287, 288, 289, 290, 291, 292, 293, 294, 295, 299, 300, 303, 304, 305, 450, 451, 454, 455, 471, 478, 479, 484, 487, 501, 505, 506, 507, 510, 515, 530, 532, 538, 541, 543, 570, 571, 572, 577, 578, 584, 591, 594, 613, 621
Lifshitz, I. M., 541
Lillie, C. F., 128
Linde, A. D., 583, 656, 657
Lindquist, R. W., 45
Ljubinov, V. A., 413
Lobachevsky, 634
Longair, M. S., 71, 72, 75, 77, 125, 129, 194, 199, 222, 377
Lorentz, H. A., 634
Low, F. J., 127
Lucchin, S., 270
Lukash, V. N., 402, 479, 539, 541, 545, 546, 550, 552, 554, 555, 559, 563, 564, 587, 593
Lynden-Bell, D., 412, 479
Lyubushin, A. A., 586

MacCallum, M. A. H., 532, 538, 550
Macdonald, D., ii, xvi, 608
Mach, E., 99, 582, 625, 626, 628, 629, 631, 632, 633, 648, 649, 654
Magalinskiy, V. B., 264
Mamaev, S. G., 593
Markov, M. A., 112, 113, 593
Marochnik, L. S., 264
Martinov, V. K., 173
Marx, G., 413, 414, 415
Masstepanenko, V. M., 593
Mathewson, D. S., 412
Matsuda, T., 394
Mattig, W., 35, 56, 59
Matzner, R., 162, 523, 541, 653
Mazets, E. P., 225, 233
McClelland, J., 413
McCone, A. I., 264
McCray, R., 233
McCrea, W. H., 3, 14
McKellar, A., 132
McVittie, G. C., 35, 47, 61, 66
Migulin, V. V., 173
Mikhailovskiy, A. B., 264
Millikan, R. A., 171
Milne, E. A., 3, 21, 258, 444, 486, 551, 552, 554, 561, 564,
Mishustin, I. N., 386
Misner, C. W., 477, 479, 495, 511, 519, 520, 521, 523, 541, 544, 545, 547, 582, 610, 612, 632, 669
Monin, A. S., 311, 336
Morganstern, R. E., 652, 653
Morpurgo, G., 173
Morrison, P., 129
Muehlner, D. J., 132, 133
Myshkis, A. D., 331, 395, 437

Nariai, H., 355, 651, 663
Narlikar, J. V., 481, 485, 624, 625, 667, 668
Narozhny, N. B., 598
Ne'eman, Y., 400, 642
Newton, I., 633, 670
Nikishov, A. I., 598
Nordsiek, K. H., 408
Northover, K. J. E., 225
Novikov, I. D., iii, iv, xv, xvii, xviii, 95, 109, 113, 120, 124, 127, 278, 294, 295, 304, 331, 342, 352, 357, 391, 392, 400, 401, 402, 412, 429, 460, 479, 501, 506, 510, 511, 521, 523, 535, 539, 541, 545, 546, 550, 552, 554, 555, 559, 560, 562, 563, 564, 572, 609, 631, 637, 638, 642, 643, 644
Novikov, S. P., 541, 549, 555
Novokreshchenova, S. I., 372
Nutku, Y., 652

O'Dell, S. L., 80, 101, 308
Oke, J. B., 47
Okun', L. B., 168, 173, 518, 583, 613, 615, 619, 657
Omnès, R., 141, 142, 144, 145, 146, 149, 168, 353, 427, 582
Oort, J. H., 394, 411
Ostriker, J. P., 412
Ozernoy, L. M., 193, 207, 229, 253, 295, 337, 355, 356, 357, 358, 377, 394, 396, 398, 399, 403, 412

Paal, G., 637
Paradoksov, P., 249, 588
Parijsky, Yu. N., 119, 120, 225, 445, 446
Parker, E. N., 388
Parker, L., 593, 606, 665
Partridge, R. B., 83, 128, 225, 445, 559
Pasierb, E., 414
Pati, J. S., 613, 614, 615
Pauli, W., 14, 35, 606
Peach, V. J., 80, 85
Pecker, J. C., 98
Peebles, P. J. E., 3, 83, 124, 128, 129, 156, 163, 175, 195, 270, 274, 275, 278, 283, 317, 322, 326, 337, 341, 351, 352, 357, 380, 391, 392, 393, 396, 406, 410, 412, 427, 430, 442, 457, 570
Peierls, S. E., 123
Penrose, R., 86, 570, 571, 575, 576, 603, 607, 653, 663, 666
Penzias, A. A., 124
Peres, A., 15
Perrenod, S. C., 219, 220, 224, 225
Peterson, B. A., 221, 370
Petrosian, V., 103, 106, 631
Petrov, 583
Petrov, A. Z., 286, 530, 532, 534, 539
Phillips, R. J. H., 414
Pikel'ner, S. B., 83, 168, 173, 196, 346, 406
Pitaevsky, L. P., 590, 594
Planck, M., 569, 585
Podurets, M. A., 462
Poincaré, H., 634
Pokrovsky, G. I., 174
Pomeranchuk, I. Ya., 149, 600
Pontecorvo, B. M., 15, 156, 414
Pooley, G. G., 72
Popov, V. N., 588, 598
Press, W. H., 340, 341, 342, 343, 607
Prilutsky, O. F., 229

Rädler, K. H., 388
Ramsey, W., 221
Rayzer, Yu. P., 430
Readhead, A. C. S., 78
Reber, G. R., 127
Redmount, I. H., ii
Rees, M. J., 75, 342, 355, 393, 398, 468
Reeves, H., 184, 186, 187
Refsdal, S., 106
Reines, E., 413, 414
Reinhardt, M., 385, 410, 495
Riemann, G. F. B., 634
Roach, F. E., 127, 128, 222
Roberts, M. S., 98, 407, 410, 495
Robertson, H. P., 61
Roeder, R. C., 86
Roll, P. G., 124
Romm, M., 139
Rood, H., 83, 409, 410
Rosen, G., 493, 495
Rozental', I. L., 229
Rozhansky, V., 352
Ruban, V. A., 491, 535, 648, 653
Rudnitsky, G. M., 372
Rutherford, E., 121
Ruzmaykin, A. A., 378, 385, 386, 388, 389, 663
Ruzmaykina, T. V., 392, 663
Ryaben'kii, V. S., 332
Ryan, M., 610, 653
Ryle, M., 71, 72

Saar, E., 405, 408, 412
Sachs, R. K., 434, 438, 446
Sagdeev, R. Z., 430
Sakharov, A. D., 143, 156, 280, 282, 283, 284, 390, 456, 613, 615, 668, 669
Salam, A., 613, 614, 615, 627
Salpeter, E. E., 103, 407, 628, 631
Sandage, A. R., 35, 44, 47, 79, 81, 83, 84, 85, 406, 442, 444, 468, 469
Sargent, W. L. W., 184, 186, 410
Sato, H., 394
Savedoff, M. P., 251

Schechter, P., 340, 341, 342, 343
Schiffer, J. P., 173
Schluter, A., 387
Schramm, D. N., 83, 175, 184, 186
Schücking, E., 479, 487, 526, 530, 538, 550
Sciama, D. W., 583
Searle, L., 184, 186
Seldner, M., 410
Setti, G., 80
Shandarin, S. F., 332, 342, 347, 365, 369, 374, 399
Shapiro, I., 651, 652
Shavokhina, N. S., 593
Shepley, L. C., 541
Shikin, I. S., 484
Shivanandan, K., 132
Shklovsky, I. S., 103, 631
Sholomitsky, G. B., 72
Shvartsman, V. F., 180, 464, 637, 640
Silk, J., 75, 270, 273, 280, 342, 355, 390, 397, 409, 412, 430, 445, 446, 447
Sil'vestrov, V. V., 264
Singwi, K. S., 123
Slysh, V. I., 86, 89
Smith, L. L., 127, 128, 222
Smorodinsky, Ya. A., 15, 181, 232, 625
Sobel, H. W., 414
Sofue, Y., 385
Sokholov, I. A., 352
Sokolov, D. D., 637, 640
Spinrad, H., 85
Spitzer, L., 198, 203, 346
Sretenskiy, L. N., 480, 482
Stabell, R., 106
Stanyukovich, K. P., 331, 593, 625, 629
Starobinsky, A. A., 303, 401, 402, 539, 587, 593, 603, 606, 607, 637, 641, 643, 644, 664, 666
Stecker, F. W., 233
Steenbeck, M., 388
Steigman, G., iii, iv, xvi, xix, 220, 583, 617
Stevens, C. M., 173
Strauss, E., 326
Struminsky, B. V., 144
Suchkov, A. A., 264
Sunyaev, R. A., xviii, 104, 125, 127, 128, 129, 179, 189, 193, 194, 195, 199, 207, 208, 210, 213, 214, 215, 221, 222, 223, 225, 226, 342, 358, 361, 370, 372, 374, 376, 389, 414, 415, 428, 445
Süveges, M., 637, 638
Syrovatsky, S. I., 229
Szalay, A. S., 413, 414, 415
Szekeres, P., 103
Szkody, P., 84

Tagirov, E. A., 603, 663
Takeda, H., 394
Talbot, R. J., 84
Tammann, G. A., 83, 444
Tarter, J., 412
Taub, A., 530
Tavkhelidze, A. N., 144
Taylor, B. Z., 85
Taylor, R., 526
Teukolsky, S. A., 607
Thaddeus, P., 132
Thiel, M., 385, 495
t'Hooft, G., 613, 670
Thorne, K. S., ii, xvi, 110, 495, 526, 610, 632, 644, 669
Tifft, W. G., 410
Tinsley, B. M., 85
Tiunov, E. A., 378
Tolman, R. C., 14, 660, 661
Tomasko, M. G., 346
Tomita, K., 355, 394, 396
Traat, R., 408
Trautman, A., 577
Tremaine, S., 413, 415
Truran, J. W., 184
Tucker, W. H., 127
Turkevich, A., 123, 174
Turley, R., 654
Turner, E. L., 412
Tyson, J. A., 462

Udin, V. M., 555

Vainshtein, S. I., 173, 378, 388, 389
Vajk, J. P., 495
van den Bergh, S., 83, 408
Veltman, M., 670
Verschuur, G. L., 412
Vette, J. I., 226, 233
Vigier, J. P., 98
Vila, S., 251
Vladimirov, Yu., 161
Vlasov, A. A., 275
von Laue, M., 35
von Weizsäcker, C. F., 355, 394
Vorontsov-Vel'yaminov, B. A., 406, 409, 410

Wagoner, R. V., 175, 178, 462
Wakano, M., 110
Warren, I. B., 541
Weber, J., 157, 239, 453, 462, 468, 652
Weinberg, S., 15, 270, 583, 627
Weiss, R., 132, 133
Wellin, A., xv
Wells, H. G., 670
Weymann, R., 192, 208, 221
Wheeler, J. A., 35, 42, 110, 582, 583, 586, 603, 610, 611, 632, 645, 669, 672
Whisnant, K., 413

Whitehurst, R. N., 407
Whitford, A. E., 47
Whittaker, E. T., 14
Wild, P., 408
Wilkinson, D. T., 124
Wilson, M., ii, xvi, 419
Wilson, R. W., 124
Winter, A. J. B., 412
Winter, K., 412, 414
Wolf, J. A., 132, 638
Wolfe, A. M., 434, 438, 446
Woltjer, L., 80
Wright, M. C. H., 412
Wroe, D., 123

Yaglom, A. M., 311, 336
Yahil, A., 412
Yakubov, V. M., 175
Yang, C. N., 140
Yu, J. T., 270, 274, 283, 442

Zakharov, V. E., 430
Zatsepin, G. T., 232
Zel'dovich, Ya. B., iii, iv, xv, xvii, xviii, xix, 3, 15, 66, 77, 86, 87, 88, 109, 120, 123, 141, 143, 156, 168, 173, 175, 181, 195, 199, 206, 208, 210, 213, 214, 215, 225, 232, 252, 262, 264, 278, 295, 303, 304, 326, 327, 329, 331, 339, 341, 342, 346, 352, 355, 357, 358, 361, 370, 372, 374, 376, 377, 389, 391, 392, 395, 401, 413, 414, 415, 424, 428, 429, 430, 432, 437, 445, 460, 462, 463, 479, 482, 495, 496, 497, 501, 511, 521, 523, 535, 540, 555, 583, 587, 588, 590, 593, 598, 602, 603, 606, 607, 609, 613, 615, 616, 617, 625, 631, 637, 638, 643, 644, 654, 655, 657, 664
Zel'manov, A. A., 21, 296, 378, 485, 530, 535
Zweig, G., 143
Zwicky, F., 73, 88, 408, 410

SUBJECT INDEX

Absolute magnitude, 56
Abundances of light elements, 185. *See also* Nucleosynthesis
Accelerated reference frame in flat spacetime, 572–574
Acoustic waves. *See* Sound waves
Action at a distance, 672
Adiabatic invariance, breakdown of, related to particle creation, 505, 588–589, 604–605, 663–665
Adiabatic perturbations. *See* Galaxy formation by adiabatic perturbations; Gravitational instability; Jeans theory of gravitational instability
Age of Universe: in terms of Hubble parameter and density, Newtonian analysis, 9, 18–19; in terms of Hubble parameter and density, relativistic analysis, 15; as function of redshift, 53–54; as function of redshift for small redshifts, 58
Angular diameter distance: defined, 42; as function of redshift, 42–44; as function of redshift for small redshifts, 58
Angular size as function of redshift, 42–44; for small redshifts, 59; influence of inhomogeneity of Universe on, 86–89
Anisotropy of Universe: places limits on its topology, 638. *See also* Isotropization of cosmological models; Isotropy of Universe; Homogeneous, anisotropic cosmological models; Homogeneous, anisotropic cosmological models of type I; Relic radiation, isotropy of
Antenna temperature, 129
Anthropic principle, 630
Antibaryons. *See* Pairs, nucleon-antinucleon; Antimatter
Antigalaxies, 121
Antimatter: observational limits on amount of in Universe, 617; energy generation by annihilation in radiation-dominated era, 209–210; antimatter-matter ratio in early Universe, 120–121, 151–152 (*see also* Galaxy formation by entropy perturbations). *See also* Pairs; Charge asymmetry of Universe; Charge symmetric model of Universe
Antiparticle-particle pairs. *See* Pairs
Apparent magnitude: defined, 55; as function of redshift (*see* Magnitude-redshift relation)
Atomic nuclei. *See* Nuclei of atoms

Background radiation, electromagnetic: spectrum from radio to gamma-rays, 125–129. *See also* Relic radiation, electromagnetic
Background radiation, gamma-ray, 129, 226, 230–231, 233
Background radiation, optical, 222
Background radiation, radio waves: from discrete sources, 75–77; from intergalactic medium, 226–229; limits on at long wavelengths, 227. *See also* Relic radiation, electromagnetic
Background radiation, ultraviolet, 222–223
Background radiation, X-ray, 194, 219–220, 230
Baryon conservation, violations of: by elementary particle processes, 613–616; by C-field processes, 623–625; by black hole processes, 615–616
Baryons, interaction of, 141–143
Beryllium: abundance of, observed, 185; nucleosynthesis by cosmic rays, 187; nucleosynthesis in early universe, 178
Bianchi classification of cosmologies. *See* Homogeneous, anisotropic cosmological models
Blackbody radiation. *See* Planck spectrum
Black holes: formation from white holes, 643–644; formation by entropy perturbations, 392; formation near initial singularity of Universe (*see* Black holes, primordial); collisions of, produce gravitational waves, 462; entropy related to surface area, 607; evaporation of, 587, 607–608, 615–616; explosions of, 609; in Brans-Dicke theory of gravity, 651; rotational energy of extractable, 607; spacetime curvature in, 635; violation of baryon conservation by, 587, 615
Black holes, primordial: possibility of many tiny holes formed near initial singularity,

This index was prepared by Kip S. Thorne.

608; formation influenced by stiffness of equation of state, 585, 610; dynamics of formation, 608–609; accretion onto, 610; evaporation as source of baryon nonconservation, 587, 615; observational limits on numbers of, 609–610
Bolometric flux as function of redshift, 55; for small redshifts, 59
Boron: abundance, observed, 185; nucleosynthesis by cosmic rays, 187; nucleosynthesis in early Universe, 179
Bose-Einstein spectrum for photons, 212
Bounce of Universe: due to modification of Einstein-Hilbert Lagrangian, 663; due to quantum effects, 612, 663; due to a negative-energy field, 622; in Brans-Dicke theory, 653–654; open model with just one bounce, 661; Hoyle-Narlikar description of, using conformal invariance, 666–668; and oscillatory Universe, 659–663; difficulty of explaining relic radiation and nucleosynthesis if bounce is too late, 653; production of gravitational waves by, 665–666
Bound-free emission. *See* Recombination
Brans-Dicke theory of gravitation, 648–654; cosmological features of, 651–654
Bremsstrahlung: formulae for, 200–201; influence on electron temperature, 205–206; in radiation-dominated era, 212–214; during and after recombination, 200–201, 204, 219–220
Brightness of a distant object as function of redshift, 45–47, 55, 59. *See also* Magnitude-redshift relation
Broken symmetries, 654–658; as source of domain (mosaic) structure of Universe, 657–658

Causality and its violation, 575–576, 602, 671
Cavendish experiments, 635
Cepheid variable stars used in determining distance scale of Universe, 82
C-field, 582, 623
Chandrasekhar limit, 628
Charge asymmetry of laws of physics, 613–616, 654, 657–658
Charge asymmetry of universe, xxx, 120–121, 135, 143, 145, 146; due to charge asymmetry of laws of elementary particle physics, 613–616; due to evaporation of primordial black holes, 616
Charge-symmetric model of Universe, 163–170, 422–428; inhomogeneous variants, 168–169; interface between matter and antimatter regions, 168–169, 422–425; kinetics of particle processes in, 163–168, 422–425, photon to nucleon ratio, 164–166; freeze-out of nucleons, 164–167; initial conditions at singularity, 426–427; observational tests of, 422–428; observations imply that each galaxy cluster is matter or anti-matter, 425–426; prevention of recombination in, 226; small islands of matter and antimatter in radiation-dominated era, 423–425. *See also* Omnes theory; Galaxy formation in charge-symmetric Universe
Chemical potential for photons, 212–213
Cigarlike motions of pressure-free matter, 334, 483–484
Closed timelike lines, 575–576. *See also* Causality and its violations
Clustering laws: Press-Schechter, 341–342; self-similar, 341–342
Clusters of clusters of galaxies, observational limits on, from relic radiation, 442, 448
Clusters of galaxies, formation of, from already condensed galaxies interacting gravitationally, 339–344. *See also* Galaxy formation
Clusters of galaxies, observations, 408–412; cell or pancake structure in, 410; invisible mass in, 409, 411–412; types, 409–411
Clusters of galaxies, statistical description of distribution, 310–324
CN, interstellar, as detector for relic radiation, 131–132
Cold Universe, 123–124; and quarks, 173; creation of entropy in, by damping of perturbations, 616–622
Collisionless gas: evolution of long-wavelength perturbations in, 339–345; gravitational instability of, 262–264
Coma cluster, 225, 371, 409–411
Combustion, theory of, 424
Comptonization parameter y, 216–220
Compton scattering: formulae for, 200–203, 207–209; quantum electrodynamics of, 595; statistical description of, Kompaneets equation, 208–209; in late stages of expanding Universe, 200–204; influence on electron temperature, 206–207; influence on spectrum of relic radiation, 192–193, 215–219, 227–228; produces Bose-Einstein photon distribution in radiation-dominated era, 210–213
Compton scattering, induced, classical theory of, 206–207
Conformal invariance: defined, 591–592; and Hoyle-Narlikar attempt to describe bounce of Universe, 666–668; influence on particle creation near cosmological singularity, 591–594; of Maxwell's equa-

tions, 603; of equation for massive scalar field, 603; of wave equation for neutrino field, 603; violated by gravitational interaction of massive particles, 608; violated by gravitational waves, 621
Conformally flat metric, 592
Continuous creation of matter, 582, 622–625. *See also* Steady state cosmology
Correlation function, 315–317
Cosmic microwave radiation. *See* Relic radiation, electromagnetic
Cosmic ray electrons: energy loss by interaction with relic radiation and starlight, 229–231; energy loss by synchrotron radiation, 229–230
Cosmic rays: in brief, 194–195; highest observed energies, 231; interaction with relic neutrinos, 232; interaction with relic radiation, 229–233; possible local origin of most energetic, 232; showers in Earth atmosphere, 231
Cosmological constant, xxii, 99–114, 585, 631; history of early research, 99–101; and stress-energy tensor of vacuum, 102, 656–657; influence on expansion of Universe, 102–108; explanation for 1.95 redshifts of quasars, 101, 103–104, 631; in combination with finite rest masses of neutrinos, 414; large in early Universe if vacuum displays broken symmetry, 656; observational limit on, 99
Cosmological principle, 623
Cosmological redshift. *See* Redshift, cosmological
Coulomb scattering, cooling of plasma by, in late stages of expanding Universe, 203–205
Covariance function, 317
CP violation, 654, 657–658. *See also* Charge asymmetry of laws of physics
Crab pulsar as possible source of the most energetic cosmic rays, 232
Critical density, xxiv; Newtonian analysis, 6–7; as determinant of future evolution of Universe, 16–17; related to curvature of Friedmann models, 25
Crystal as example of broken symmetry, 655
Cutoff in radio source evolution, 72–73
Cutoff momentum in quantum electrodynamics, 600; related to gravitational effects, 627. *See also* Renormalization

Damping mass, 274
Deceleration parameter: Newtonian analysis, 7; in steady state cosmology, 625; observational determination of, 84–88; related to closure of Universe, 61
Deformation of reference systems, 296–297
Deformation tensor. *See* Velocity of deformation tensor
Delbruck scattering, 601
Density of Universe, 194; in brief, xxv–xxvii, 8; best current estimate of, 412; inferred from existence of neutral hydrogen halos of galaxies, 223–224; inferred from observed deuterium abundance, 188; inferred from peculiar velocities of galaxies, 443–444; inferred from stability of clusters of galaxies, 411–412; inferred from observed age of Universe, 15–16; possible forms of "missing mass," 411–415
Density parameter: defined, 7; as function of redshift, 54; redefined when pressure is not negligible, 60
Deuterium: abundance observed, 185–187; destruction in stars, 186; nucleosynthesis in anisotropic cosmological models, 527–529; nucleosynthesis in expanding Universe, 178–180; nucleosynthesis in supernovae and supermassive stars, 186
Differential forms, 533
Dirac's theory of large numbers, 583, 626–629
Distance scale of Universe: observational determination of, 81–83; revisions of, as observations have improved, 100
Domain structure of Universe resulting from broken symmetry of vacuum, 657–658
Dynamo effect, 387–389

Earth, peculiar velocity relative to Universe, 442–443
Einstein field equations, 22
Einstein static Universe, 103
Elasticity of spacetime, 668–670
Electric charge of closed Universe must be zero, 113
Electromagnetic radiation of Universe. *See* Background radiation, electromagnetic; Relic radiation, electromagnetic
Electron-electron scattering. *See* Coulomb scattering
Electron mass, explained by broken symmetry, 656, 658
Electron-positron pairs. *See* Pairs, electron-positron
Electron scattering of photons. *See* Compton scattering
Electron-scattering optical depth as function of redshift, 74
Elementary length of 10^{-17} cm at which laws of physics might change, 569

Elementary particles of unknown type: possible roles in early Universe, 157, 180–181; with Planck mass, 593
Elementary particles, spectrum of, 139–141; Hagedorn theory, 147–148
Ellipsoidal body, gravitational collapse of, 480–484
Elvert formula, 200
Entropy: effect of gravity on, 645–647; increase of, related to arrow of time, 670–672; of a black hole, 607; of relic radiation, 155–156
Entropy barrier to nucleosynthesis, 178
Entropy of Universe: origins of, in brief, 355; due to charge asymmetry of laws of elementary particle physics, 614; due to damping of initial perturbations, 616–622; due to dissipation of short-wavelength acoustic waves, 582, 616–619; due to evaporation of primordial black holes, 615, 610; due to neutrino interaction with anisotropic expansion, 516–519; growth of, in oscillatory Universe, 659–662; increase of, forbids Universe to be closed in time, 638; related to gravitational fine structure constant, 631. *See also* photon to baryon ratio
Equation of state: in Hadron era, 140–142, 665; maximally stiff, 17–18, 141, 619–620; maximally stiff produces too many gravitons in early Universe, 666; influence of graviton creation on, 665
Equilibrium of electrons with non-Planck radiation field: Kompaneets equation for evolution of photons, 208–209; electron temperature, 205–207
Ergosphere of a black hole, 607
Eulerian coordinates, 329
Evolution of discrete sources of radiation: as impediment to determining cosmological parameters, 67–70, 81; radio sources, 70–78
Expansion of Universe: in brief, xxii–xxiv, xxvii; discovery of, 100; and arrow of time, 671–673; effect of pressure on, 12–14
Expansion of Universe, Newtonian evolution law: differential equation, 6–7; early stage, 9–11; late stage, 10–11; transition from early stage to late stage, 11
Expansion of Universe, relativistic evolution law, 24, 37; curvature, Hubble parameter, and density as functions of time, 24; differential equations, 14–15, 22–23; early stage, 17–18; future evolution depends on density, 16–17; with nonzero cosmological constant, 102–108
Ferromagnet as example of broken symmetry, 655
Feynman diagrams, 149–150, 594–601
Feynman path integrals for evolution of Universe, 610–612
Fine structure constant, electromagnetic, what if it were of order unity or larger?, 626
Fine structure constant, gravitational, 159, 585, 618, 626–631
Foamlike structure of spacetime at Planck length, 645
Free-free emission. *See* Bremsstrahlung
Friedmann cosmological equations, 22–23
Friedmann cosmological models, 21–32; history of early research on, xxxv, 99–101; line element, 36, 21; conformal flatness of, 592; topology of, is indeterminate, 635–639; curvature parameter K, 21; scale factor, 21–22; radius of curvature of space, 21, 27; differential equations for evolution, 22–23; evolution of, 23–24, 37; observational determination of parameters, xxiv–xxv, 66–70; observational features of, 33–989 (*see also* under specific features, e.g. Redshift); scale factor defined, 21–22; topology of, is indeterminate, 635–639; with nonzero cosmological constant, 102–108
Friedmann cosmological models, closed: geometry of, 23, 25–27; volume of, 27; evolution of, 24–25; total number of particles in, 40
Friedmann cosmological models, flat: geometry of, 32; 18 possible topologies of, 638; evolution of, 24, 32; density equal to critical, 31–32
Friedmann cosmological models, open: geometry of, 28–29; infinite number of possible topologies of, 638; evolution of, 24, 29–31; low-density limit, 29–31
Fundamental constants of physics, time variation of, 633, 650

Galaxies, clusters of. *See* Clusters of galaxies; Intracluster gas
Galaxies, distribution of, described statistically, 310–324
Galaxies, first generation of, 370
Galaxies, halos of, 412–413; formation by fragmentation of pancake protoclusters, 376–377
Galaxies, magnetic fields of, xxxi–xxxii, 351, 384–388, 495
Galaxies, nuclei of, formation by fragmentation of pancake protogalaxies, 376

Galaxies, observations of, 405–412; diameters, 407; luminosity function, 406, masses, 406–407; rotation, 406–408; types of, and their distribution, 408; velocity dispersions, 408; neutral-hydrogen gas near, 412; neutral-hydrogen bridges between, 223–224; neutral-hydrogen halos, 223–224

Galaxies, peculiar velocities of, 442–444; produced by gravitational waves, 467–469

Galaxies, rotation of: observations, 406–408; origin, xxxi, 351, 377–384

Galaxy formation: for the mathematical analysis of the evolution of perturbations, *see* Gravitational instability; for applications of that analysis to galaxy formation, *see* this and following entries

Galaxy formation, 351–418; in brief, xxix–xxxi, 237–238; comparison of theories of, 350–358, 399–405; tests of theories by searches for 21 cm radio waves from condensations, 371–372; initial conditions at singularity, 292, 294–299, 298–299, 301–305, 460, 593, 618, 640; equipartition hypothesis for initial conditions, 460; magnitude of perturbations required at beginning of Universe, 294; requires same initial perturbation amplitude as for entropy production, 616–619; magnitude of perturbations required as function of redshift, 328; magnitude of perturbations required at recombination, 260–262. *See also* the next six major entries

Galaxy formation by adiabatic perturbations, 350–351, 358–389; in brief, 294; mathematical theory of adiabatic perturbations, 247–250, 270–274, 290–294, 359; initial conditions at singularity, 334–335, 404–405, 584–585 (*see also* Initial conditions at singularity); overview of evolution of perturbations, 359–360; phase relation between density and velocity perturbations, 438–439; influence of gravitational waves on, 463–465; absence of vorticity at recombination, 378; formation of adiabatic perturbations by vortex perturbations after recombination, 395–396, 404, 416–418; formation of pancakes, extrapolated linear theory, 329–335, 347–349, 360–366; absence of cigar motions in pancakes, 334; shock waves in pancakes, 360–366; cooling of pancakes, 366–372; fragmentation of pancakes, 372–377; statistical description of evolution of pancakes, 372–374; observations of pancake structure in galaxy distributions, 410, observable radiation from pancake protoclusters, 370–372; test of theory by searches for 21 cm radio waves from pancakes, 371–372; implications for relic radiation, 428–432, 445–448; peculiar motions of galaxies and clusters, 421, 438–440; origin of galactic rotation, 351, 377–384; origin of galactic magnetic fields, 351

Galaxy formation by entropy perturbations, 351–352, 354, 389–393; mathematical analysis of entropy perturbations, 250, 276–278, 302–303; assessment of theory, 405; marriage with theory of adiabatic perturbations, 392–393; initial spectrum of perturbations at singularity, 389–390; nonlinear stage of evolution, 390–391; formation of antimatter regions in Universe, 354–355; formation of globular clusters, 352, 391–393; formation of massive black holes, 392; formation of supermassive stars, 391; reionization of intergalactic medium by supermassive stars, 391; clustering of small condensations to make large ones, 339–344; limits on spectrum of perturbations from observations of relic radiation, 427–428

Galaxy formation by explosion of a superdense body (Ambartsumyan theory), xxxiii, 400–401

Galaxy formation by turbulence. *See* Galaxy formation by vortex perturbations

Galaxy formation by vortex perturbations, 355–358, 395–399, 401–405; mathematical theory of vortex perturbations, 250–251, 278–280, 295–299; free parameters of theory, 397, 405; initial conditions at singularity, 401–402, 405, 593; formation of adiabatic perturbations from vortex perturbations, 395–396, 404, 416–418; nonlinear regime of evolution, 416–418; supersonic turbulence after recombination, 394; weak-turbulence variant of theory, 395–396; pancake formation and fragmentation into galaxies, 396; clustering of small objects to form larger objects, 398; explanation of rotation of galaxies, 356, 399; difficulties of theory, 357–358, 402–405; difficulties with isotropy of relic radiation, 403–405; difficulties in accounting for nucleosynthesis of light elements, 401–402

Galaxy formation in charge-symmetric Universe, 352–355; observational contradictions of theory, 422–428

Gamma-rays: background, 129, 226, 230–231, 233; from evaporating and exploding black holes, 609–610; from antimatter annihilation in charge-symmetric Universe, 232–233
Gaussian curvature of a space of constant curvature, 21, 23, 25
General relativity: Mach's principle and, 631; anticipated by nineteenth-century research on non-Euclidean geometry, 634; experimental tests of, 632–633, 651–652, 654; homogeneous spacetimes, mathematical formalism for, 530–535; linearized approximation to, 450–451
Geodesic completeness, 573–575
Geometrical optics: fundamental principles of, 434–436; breakdown of, related to particle creation, 588; for relic radiation in anisotropic cosmological models, 557–558, 560–561. *See also* Adiabatic invariance
Geometry of Universe (finite versus infinite), observational evidence for, xxiv–xxv. *See also* Critical density; Density of Universe; Friedmann cosmological models; Topology of Universe
Globular clusters, formation by entropy perturbations, 278, 352, 391–393, 427
Gravitational collapse: of an ellipsoidal body, 480–484; effects of quantum gravity on, 590–591; particle creation near singularity at endpoint of, 591; singularities in (*see* Singularities). *See also* Black holes; Gravitational instability; Galaxy formation
Gravitational constant: magnitude of, determined by elasticity of spacetime and vacuum polarization, 669–670; time variation of, 630, 651; variability of, in Brans-Dicke theory, 649–651
Gravitational energy of Universe, 635
Gravitational fine structure constant. *See* Fine structure constant, gravitational
Gravitational instability: for the mathematical analysis of the evolution of perturbations, *see* the following entries; for applications of that analysis to galaxy formation, *see* Galaxy formation
Gravitational instability, nonlinear regime, 325–349; methods of analyzing, 325; statistical description of, 335–339, 372–374; spherical perturbations, 325–328; nonlinear formation of long-wavelength perturbations compared to linear growth, 341; extrapolation of linear theory (pancake model), 329–335, 347–349, 360–366; formation of shocks, 337–339; wave-wave coupling, 336–339; fragmentation of pancakes, 372–377
Gravitational instability of expanding, homogeneous medium, Newtonian theory, 247–262; critical perturbation amplitude required for galaxy formation, 328; relation to Jeans theory, 245, 249; statistical description of perturbations, 310–324, 372–374; adiabatic perturbations, 247–250; entropy perturbations, 250, 276–278; vortex perturbations, 250–251, 278–280; spherical perturbations, 25–28, 253–255; large-scale, growing mode, 251–255, 269; formulae for evolution of perturbations, 255–259; evolution of perturbations before recombination, 265–284; effect of recombination on perturbations, 281–284; matching of perturbations before and after recombination, 280–284; growth after recombination, 259–262; damping of adiabatic perturbations by viscosity and thermal conductivity, 270–274; damping of entropy perturbations, 276–278; damping of vortex perturbations, 279–280; damping by gravitational interaction with neutrinos and gravitons, 274–276; nonlinear regime (*see* Gravitational instability, nonlinear regime)
Gravitational instability of expanding, homogeneous medium, relativistic theory, 285–309; Hawking analysis contrasted with Lifshitz analysis, 287; comoving coordinates and perturbations of metric, 285, 288–289; coordinate transformations, 286–289; hyperspherical harmonics in analysis of, 305–306; equations governing perturbations, 289; solutions of perturbation equations when pressure vanishes, 294, 299; decomposition of perturbations into scalar, vector, and tensor modes, 289–291; adiabatic (scalar) perturbations, 290–294, 359; entropy (scalar) perturbations, 302–303; vortex (vector, rotational) perturbations, 295–299, 593; tensor (gravitational wave) perturbations, 299–302; unphysical modes, 287, 289; quasi-isotropic solution of perturbation equations, 303–305; statistical homogeneity and isotropy of perturbations, 307–309; perturbations of size comparable to radius of Universe, 305; initial conditions at singularity, 292, 294–295, 298–299, 301–305, 399, 460, 593, 618, 640 (*see also* Initial conditions at singularity); hypothesis of uniform distribution of perturbations, 304–305.

See also Homogeneous anisotropic cosmological models of type I, perturbations of
Gravitational instability of stationary, homogeneous medium, statistical mechanics explanation of, 645–647. *See also* Jeans theory of gravitational instability
Gravitational mass defect, 112
Gravitational radiation. *See* Gravitational waves
Gravitational waves, theory of: in brief, 239; general relativistic theory of, 450–454; speed of propagation, 452; in Brans-Dicke theory, 652; quantization of, 157–158, 587–588, 663; coupling to elementary particles, 159–161, 458; stress-energy tensor for, 451–452; stochastic superposition of, 452–453, 455–456, 467–468; generation by ellipsoidal matter motion, 489; interaction with an ideal fluid, 300; interaction with electromagnetic fields, 470–474; damping by a viscous fluid, 300–301; deformation of matter by, 300. *See also* Gravitational waves, cosmological aspects of; Gravitons
Gravitational waves, cosmological aspects of, 449–474; creation by discrete bodies (pulsars, quasars, black holes, etc.), 461–462; creation in bounce of Universe, 665–666; creation near cosmological singularity, 160–163, 455–459, 663–666; amplification near cosmological singularity, 621, 663–666; equilibration with other forms of energy at Planck time, 160–163, 455–459; creation by tensor perturbations, 454–456; evolution as Universe expands, 158, 299–300, 453–454; predicted spectrum in Universe today, 163, 457–461, 515–516, 519, 621; in Kasner universe, 504–506; in Bianchi type VII universe, 555; minimum amplitude for significant cosmological effects, 453; influence on primordial nucleosynthesis, 464–465; influence on galaxy formation, 463–465; influence on isotropy of relic radiation, 465–467; influence on peculiar velocities of galaxies, 467–470; observational search for, 453, 462, 472–474, 652; observational limits on spectrum of, 463–470
Gravitons: obtained by quantizing gravitational waves, 157–158, 587–588, 663; related to adiabatic invariant for classical gravitational waves, 454; rest mass zero, 158, 663; cross section for interaction with elementary particles, 160–161, 458; decoupling from matter in early Universe, 515; evolution of in anisotropic cosmological models, 511–516; gravitational interaction with sound waves, 274–276; influence on nucleosynthesis in early Universe, 180–181; interaction with background spacetime in quantum theory of gravity, 669–670; mathematics of creation by strong gravitational field, 664–665. *See also* Gravitational waves
Gravity, quantum theory of. *See* Quantum gravity
Group of motions, 532

Hadron era, 136, 139–152;
Hagedorn theory, 141, 147–150, 162, 582; contrasted with standard theory, 137; cosmological consequences of, 583–585; influence on number of free quarks and partons left over from big bang, 583
HA models. *See* Homogeneous, anisotropic cosmological models
Heat death of Universe. *See* Thermal death of Universe
Heckmann-Schucking solution, 487–488, 526
Helium: observed abundance, 184–185; nucleosynthesis in early Universe, 178–182; nucleosynthesis in anisotropic cosmological models, 527–529; nucleosynthesis in stars, 184, 186–187
Hierarchical Universe, contradicted by observed isotropy of relic radiation, 441–442
History of cosmological research, xxxi, xxxv, 99–101, 121–125, 132, 174–175, 240, 285, 571, 631, 633–635, 670–672
Homogeneity: mathematical definitions of, 531–533, 535–537; differential homogeneity, 535–537; group homogeneity, 532–533; physical definition of, 530–531, 533
Homogeneity and isotropy of Universe, problem of, in light of existence of horizons, 477–478, 540–541
Homogeneity of Universe: in brief, xxxi–xxxii, embodied in Friedmann cosmological models, 20–21; preserved by Hubble expansion, 4–5; limits on, from isotropy of relic radiation, 421, 442; limits on, from observations of galaxy distributions, 410; statistical formulation in inhomogeneous models, 307–309
Homogeneous, anisotropic cosmological models, 530–565; definition of homogeneity, 530–533, 535–537; Bianchi classification of, 533–535; structure constants for, 533–534; metric for, 533; Einstein field equations for, 538; con-

servation of topology of, 639; conservation of deformation anisotropy in radiation-dominated era, 560, 563–564; dynamics near singularity, 537–548; dynamics near singularity for type V, 552; dynamics near singularity for type VII, 555; isotropization of, 548–555; isotropization of type V, 550–552; isotropization of type VII, 553–554; isotropization of type IX, 552–553; limits on, from observed isotropy of relic radiation, 564–565; degenerate cases, 540; special model with topology $S2 \times E1$, 535; type I, 534, 537, 550, 563 (*see also* Homogeneous, anisotropic cosmological models of type I); type II, 534, 539; type III, 534, 537, 539; type IV, 534, 539; type V, 534, 539, 550, 560–563; type VIh, 534, 539; type VIo, 534, 539; type VII, 534, 537, 550, 553–555, 560, 563–565; type VIIh, 534, 535, 539; type VIIo, 534, 539; type VIII, 534, 540, 577; type IX, 534, 535, 540, 550, 552–553, 560, 563–564, 577, 580, 612 (*see also* Mixmaster cosmological model)

Homogeneous, anisotropic cosmological models of type I, 477–565; motivations for studying, 474–479; Newtonian, 480–485; Newtonian and relativistic compared, 479; analytic solutions of Einstein equations, 485–487, 493, 497–499; Kasner solution, 485–489; Transition from Kasner stage to matter-dominated stage, 490–491; Heckmann-Schucking solution, 487–488; solutions with magnetic field, 493–500; solutions with magnetic field and matter, 496–500; effect of anisotropic stress-energy on, 491–495; effect of viscosity on, 519–521; effect of free-streaming particles on, 494, 513–516, 521–523; isotropization of, 482–483, 490–491, 493, 497; isotropization in presence of free-streaming particles, 513–515; isotropization in presence of interacting neutrinos, 517–519; isotropization in presence of magnetic fields and free-streaming particles, 556–557; isotropization in presence of viscosity, 519–521; thermodynamic equilibrium in earliest stages of, 524–525; gravitons in, 511–516; neutrinos in, interactions with each other and electrons and positrons, 516–519; neutrinos in, decoupling from matter, 516; neutrinos in, free-streaming approximation, 511–516; neutrinos in, present-day neutrino energies, 511; nucleosynthesis in, 523–529; observational limits on, from primordial element abundances, 528–529; predicted anisotropy of relic radiation, 555–559 (*see also* Relic radiation, isotropy of); perturbations of (*see* Homogeneous, anisotropic cosmological models of type I with perturbations)

Homogeneous, anisotropic cosmological models of type I with perturbations, 501–510; Newtonian, 502–503; initial conditions at singularity, 510; perturbations of dust in Kasner solution, 504; gravitational waves in Kasner solution, 504–506; homogeneous perturbations of matter motions in Kasner solution, 506–510; creation of spatial curvature by matter motions, 509–510. *See also* Particle creation by strong cosmological gravitational fields

Homogeneous, isotropic cosmological models. *See* Friedmann cosmological models

Horizon, particle, 36–41, 540, 630; and topological structure of Universe, 639–640

Hot model of Universe: in brief, xxvii–xxix, 117–125; epochs in evolution of, 134–136; evolution of energy densities in, 158; history of ideas on, 121–125; neutrinos in, 153–157; thermodynamic equilibrium in, 136–138

H II regions used in determining distance scale of Universe, 83

Hubble constant. *See* Hubble parameter

Hubble diagram. *See* redshift-magnitude relation

Hubble parameter: defined, 4; evolution of, 52–53; observational determination of, 8, 81–84, 442; uncertainty in, affects affects our knowledge of closure of Universe, 194, 224

Hyades cluster, 81

Hydrodynamics: equations of, 240; Lagrangian formulation of, 332; example of broken symmetry in, 655; vorticity and its evolution, 378–380; shock waves (*see* Shock waves); sound waves (*see* Sound waves); turbulence (*see* Turbulence)

Hydrogen, H^-: formation of, in expanding Universe, 199

Hydrogen, ionized intergalactic, observational limits on, 193–194

Hydrogen, molecular, formation of, in expanding Universe, 199

Hydrogen, neutral atomic: absence of, in clusters of galaxies, 224; absence of, in intergalactic space, 193, 221–222, 370; around galaxies and bridging galaxies, 223–224, 412; epoch of, after recombination, 226, 229; studied by 21 cm radiation, 223